Mathematical Biosciences

an international journal

Volume 66 Number 1 September 1983

DANIEL S. LEVINE (Houston, TX)
Neural Population Modeling and Psychology: A Review 1

L. BASS AND A. J. BRACKEN (Brisbane, Australia)
The Flux-Ratio Equation under Nonstationary Boundary Conditions 87

EMMANUEL T. RAKITZIS (Athens, Greece)
The Concentration of Modified Protein Species as an Index of Protein Modification Cooperativity 93

WILLIAM H. E. DAY (St. John's, Newfoundland, Canada)
The Role of Complexity in Comparing Classifications 97

GEORGE MARKOWSKY (Yorktown Heights, NY)
Necessary and Sufficient Conditions for a Phenotype System to Have a Factor-Union Representation 115

J. EISENFELD (Arlington, TX) AND S. M. GRUNDY (Dallas, TX)
Structural Identification of Large Systems by Reduction to Subsystems: VLDL Triglycerides 129

ERRATA 151

BOOK REVIEW

CATHERINE A. MACKEN
An Introduction to Genstat (N. Alvey, N. Galwey, and P. Lane) 153

BOOKS RECEIVED 155

Elsevier Biomedical

Papers to Appear in Forthcoming Numbers

EBERHARD O. VOIT AND GEORG DICK
Growth of Cell Populations with Arbitrarily Distributed Cycle Durations II. Extended Model for Correlated Cycle Durations of Mother and Daughter Cells

ALEX S. PAPADOPOULOS
A Stochastic Model for Bod and Do with Random Initial Conditions and Distance-dependent Velocity

HANS-ROLF GREGORIUS
Efficiency of Random Pollination and Optimal Sex Ratio

CHIN LONG CHIANG
The Theory of Multistage Carcinogenesis

GEORGE MARKOWSKY AND ANDREW WOHLGEMUTH
Identifying Antigens and Antibodies in Serology

Mathematical Biosciences is published 10 times a year (monthly except January and March) by Elsevier Science Publishing Co., Inc. Subscription price: $325.00. For air mail delivery inside the United States and Canada, please add $22.00. For surface mail delivery outside the United States, please add $16.00. For air mail delivery to Europe and Mexico, please add $44.00. For air mail delivery to the rest of the world, please add $56.00. Single copy and back volume information available on request. Claims for missing issues can be honored only up to three months for domestic addresses and six months for foreign addresses. Duplicate copies will not be mailed to replace ones lost through failure to notify Elsevier of change of address.

Postmaster: Please send address changes to Elsevier Science Publishing Co., Inc., 52 Vanderbilt Ave., New York, NY 10017.

Please direct orders for this journal, change of address, and claims for missing issues to: Journals Fulfillment Department, Elsevier Science Publishing Co., Inc., 52 Vanderbilt Ave., New York, NY 10017.

Please direct inquiries regarding the placing of advertisements in this Elsevier journal to: Journals Advertising Coordinator, Elsevier Science Publishing Co., Inc., 52 Vanderbilt Ave., New York, NY 10017 (212) 867-9040.

©Elsevier Science Publishing Co., Inc., 1983. All rights reserved. Second class postage paid at New York, NY and at additional mailing offices. Manufactured in the United States.

This journal has been registered with the Copyright Clearance Center, Inc. Consent is given for copying of articles for personal or internal use, or for the personal or internal use of specific clients. This consent is given on the condition that the copier pay through the Center the per-copy fee stated in the code on the first page of each article for copying beyond that permitted by the U. S. Copyright Law. If no code appears on an article, the author has not given broad consent to copy and permission to copy must be obtained directly from the author. All articles published prior to 1978 may be copied for a per-copy fee of $1.00, also payable through the Center. This consent does not extend to other kinds of copying, such as for general distribution, resale, advertising and promotional purposes, or for creating new collective works.

Mathematical Biosciences is abstracted or indexed in Applied Mechanics Review, Biological Abstracts, Chemical Abstracts, Current Contents, Engineering Index, Excerpta Medica, International Abstracts of Biological Sciences, and Mathematical Reviews.

Neural Population Modeling and Psychology: A Review

DANIEL S. LEVINE
*Department of Mathematics, University of Houston–University Park, Houston, Texas 77004**

Received 25 March 1983

ABSTRACT

The development of neural population modeling as it relates to psychology is traced from the early 1940s to the present. The evolution of the field has been from descriptions of randomly connected neurons transmitting all-or-none signals to analyses of structured multi-level networks whose dynamics involve several different spatial and temporal scales. The cybernetic revolution of the 1940s led to the incorporation into digital neural models of such concepts as linear threshold logic, redundant computation, and information. Each of these concepts has more recently been synthesized with learning to generate a set of adaptive neural models. Concurrently, a variety of data from neurophysiology and from experimental psychology suggested models that incorporated continuous and nonlinear effects. Since the late 1960s there has been much activity in the design of rules for modifiable synapses in models of learning or conditioning. There has also been much activity in the design of lateral inhibitory networks that model sensory pattern storage. The development of models of these effects is outlined, together with that of multi-level networks that combine modifiable synapses and lateral inhibitory anatomies. These multi-level networks model such psychological effects as reinforcement, attention, coding of feature detectors, and interactions between short-term and long-term memory.

1. INTRODUCTION

The relationship between structure and function in the central nervous system poses one of the most important and difficult problems in biology. The complexity of both the connection patterns of billions of neurons and the psychophysiology of behavior strongly suggests that mathematical modeling would be useful in understanding the structure-function relationship. Yet, because of this very complexity, such modeling has proceeded in many different directions. During the last several years, some unifying themes have emerged from this diversity.

Mathematical neural modeling as we know it today began in the 1940s with the development of cybernetics and related disciplines. Many re-

*Current address: Department of Mathematics University of Texas–Arlington Arlington, Texas 76019

©Elsevier Science Publishing Co., Inc., 1983
52 Vanderbilt Ave., New York, NY 10017 0025-5564/83/$03.00

searchers expected the cybernetic approach to provide immediate, powerful insights into brain organization via analogies with digital computing machines, whose theory was being developed at the same time. Experimental findings in psychology and neurophysiology, however, revealed that the brain is organized in ways not well modeled by the computers then (or now) existing. Thus, by the late 1960s, automata theory and neural modeling had both made progress, but in directions so divergent that they had become separate fields. Automata and artificial-intelligence theorists had moved toward heuristic programs designed to handle single tasks in such areas as pattern recognition and language understanding, not necessarily using the same mechanisms as biological organisms. Concurrently, neural-network modelers were basing their models on mechanisms suggested more by neurophysiological and psychological data than by the computer analogy (although computer *simulation* has played a large role in understanding the dynamics of specific models). As the mathematics of choice for these models, discrete automata theory became less important than continuous differential equations.

This article will give an overview of neural models from the early 1940s to the present, with the aim of classifying models both by the way they explain specific phenomena (for example, conditioning, memory, or pattern discrimination) and by their mathematical structure (for example, continuous versus discrete, serial versus parallel, linear versus nonlinear, deterministic versus stochastic). Where there is widespread current agreement on the structures involved in producing a given phenomenon, the posited structures will be emphasized as a major theme of modeling. Where there is disagreement, different possible structures will be compared, together with their inherent limitations.

For the sake of brevity, this article will concentrate on population and global models and omit those dealing with single neurons. Thus, I will not discuss such important work as the Hodgkin-Huxley equations for the neuron [112], the many mathematical variations of Hodgkin-Huxley (for example, [46, 58, 60, 157]), or the models of dendritic structure by Rall (for example, [176, 177]). Some of the major single-cell models are reviewed in [187]. Assumptions about the neuron—notably its firing patterns and the modifiability of its synapses—will be reviewed as they are necessary for understanding the basis for a given population model. It is worth noting, however, that much similarity exists between the organizing principles of single neurons and those of neural networks (cf. [90, 100, 218]). In particular, a neuron itself can be treated as a "network" of membrane sites.

The organization of this review article reflects the evolution of neural population models from descriptions of randomly connected neurons transmitting all-or-none signals to analyses of structured multilevel networks whose dynamics involve several different spatial and temporal scales. This

evolution of models has tended either to follow or anticipate the evolution of knowledge in experimental neurobiology. Models that anticipated experimental findings in neurobiology were suggested either by results from experimental psychology or by neural principles that synthesized results from several fields. Thus, the exposition in this article will reflect at some times an historical experimental progression and at other times a synthetic viewpoint that organizes and predicts data from several types of experiments.

2. EARLY HISTORY AND BACKGROUND

2.1. *THE DIGITAL APPROACH*

A classical mathematical study of networks of all-or-none neurons was the 1943 article of McCulloch and Pitts [147]. Previous to that article, electrical recording had indicated that propagated neuronal signals are initiated in an all-or-none fashion [1]. Given this experimental substrate, McCulloch and Pitts generated great excitement by showing that any logical function can be duplicated by some network of neurons of an all-or-none character. That is, if a neuron were connected into a suitable network, it could be made to fire selectively in response to any given spatiotemporal array of firings of other neurons in the network. Thus, very simple dynamics could generate very complex behavior if embedded in a properly chosen network configuration.

The all-or-none neurons in the model of [147] were connected by excitatory and inhibitory pathways, according to the following rules (see Figure 1).

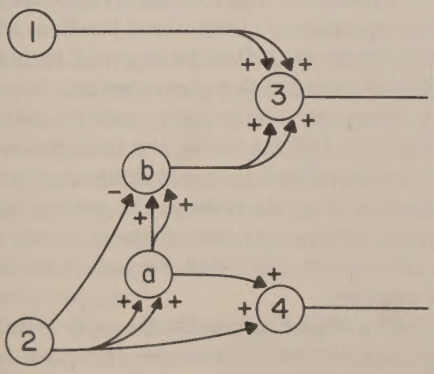

FIG. 1. Example of an all-or-none neural network. Cells labeled "1" and "2" were heat and cold receptors on skin. Heat and cold were felt when cells "3" and "4" fired, respectively. Arrows are *synapses* (junctions between cells); "+" denotes excitation, "−" inhibition. Each cell had a threshold of 2. In this network, a cold object held to skin and then removed produced a sensation of heat. (Modified from McCulloch and Pitts [147]; printed with permission from Pergamon Press, Ltd.)

Each neuron in the network obeyed a simple form of a *linear threshold* law: it fired whenever at least a given number of excitatory pathways impinging on it were active within a unit time period. All computations were carried out at discrete time intervals. If a neuron received a single inhibitory signal from an active neuron, then it was prevented from firing. The structure did not change as a function of experience; thus, the model dealt with performance, not learning.

The McCulloch-Pitts theorem inspired many researchers to strive for a more detailed understanding of the logical functions performed by the nervous system. These researchers included most notably McCulloch himself [146] and the great mathematicians von Neumann [159, 160] and Wiener [210]. This work was carried on even as the digital assumption (that neurons operate in an all-or-none manner) was being challenged. For example, von Neumann stated: "The 'digital method' used in computing may be entirely alien to the nervous system. The discrete pulses used in neural communications look indeed more like 'counting' by numeration than like a 'digitalization'" [159, p. 38].

Von Neumann was particularly interested in the self-correcting capacity of the nervous system. Two different aspects of this problem are the reliability or stability of neural computations and the adaptive nature of behavior. The former issue was approached in [159]. The basic idea of this article was that even if individual elements in an (electrical) network had a fairly high probability of malfunction, the probability of malfunction of the entire network could be made arbitrarily low if information was made to flow along parallel lines (*redundancy*). The redundancy hypothesis was supported by experiments showing retention of behavioral functions after certain brain lesions [134]. Von Neumann quantified the degree of redundancy needed to reduce the probability of error below a given amount.

Von Neumann's theory was contemporary with the theory developed by Shannon and Weaver [191, 192] for coding and transmission of information. Information theory was developed for use in designing electronic communications networks, such as telephone systems. The analogy between telephone networks and nervous systems had been suggested many years earlier by Helmholtz [128]. This insight stimulated the use of information-theoretic concepts in neural modeling.

Shannon also used a digital approach, the unit of information being defined as a binary digit, or "bit." The bit was defined as a choice between two equally probable symbols. A pattern of symbols, such as letters or numbers, was said to contain information to the extent that it was improbable.

The concept of information, like those of threshold logic and redundant computation, was originally developed for systems whose rules do not change through time. Each of these performance measures was eventually joined

with learning mechanisms to generate a set of adaptive models. Some of these suggested that learning constitutes an increase in information (see Section 6.3 of this article). Other models suggested that learning alters neural firing patterns by changing the efficacy with which signals can exceed threshold (see Sections 6.1 and 6.2). Still other models joined redundancy to learning by describing associative learning within parallel neural channels (see Chapter 8). Thus, many contemporary approaches can trace their origins to the pioneering studies of workers like McCulloch, Pitts, von Neumann, and Shannon. Despite these historical moorings, recent theories have progressively eschewed the all-or-none and random-network hypotheses, for reasons that will be developed as the article proceeds.

2.2. *SOME EARLY CONTINUOUS MODELS*

Along with the cybernetic revolution came a rapid increase in results in both experimental psychology and neurophysiology that stimulated the development of continuous neural models. One of the pioneers in this development was Rashevsky [179].

Rashevsky described how the early (1938) edition of his book used postulates based on differential equations to model such psychophysical data as reaction times and just-noticeable differences. The 1943 paper of McCulloch and Pitts had significantly altered his thinking: "the proper mathematical tool for representing the observed *discontinuous* interaction between neurons was not the differential equation but the Boolean Algebra or Logical Calculus" [179, p. 3].

It proved difficult, however, to model the observed psychophysical data using the McCulloch-Pitts postulates. This paradox was resolved when Rashevsky observed that these data reflect interactions between very large numbers of neurons. Hence, "the discontinuous laws of interaction of individual neurons lead to a sort of average continuous effect which is described by the differential equations postulated originally" [179, p. 3]. The reconciliation thus effected between continuous deterministic models and discrete random models is still in common use today. In addition, contemporary modelers often average random effects across time, thus converting potential into firing frequency.

Rashevsky made a major contribution to the incorporation of continuous effects into neural models, but he did not completely analyze the process by which all-or-none effects are averaged when many neurons interact. For example, he assumed for mathematical simplicity that the frequency of the impulses transmitted by a neuron is, on the average, a linear function of the cell's suprathreshold activity. This hypothesis also formed an important part of the Hartline-Ratliff model of the horseshoe-crab retina [108]. This is not, however, the only signaling rule that can be derived from averaging consider-

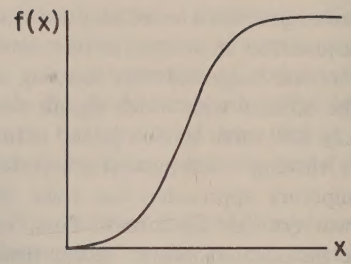

FIG. 2. Schematic graph of a sigmoid function.

ations. For example, if the firing threshold of an all-or-none neuron is described by a random variable with a Gaussian or Cauchy distribution, then the expected value of its output signal is a *sigmoid* or S-shaped function (Figure 2) rather than a linear function of suprathreshold activity [8, 90, 214].

Rashevsky recognized another limitation of some of his models: "The vast majority of the structures in this book are highly specific, in the sense that a slight variation of the structure or a possible breakdown of a part of it renders the whole mechanism inoperative" [179, p. 230]. This acute sensitivity to breakdown (in opposition to von Neumann's principle of reliability) is apparent, for example, in Rashevsky's model of classical conditioning. The model he proposed for conditioning was one in which memories were stored by the reverberation of activity within excitatory feedback loops. Consequently, any momentary perturbation that inhibited the reverberation would destroy the memory.

Contemporary neural modelers use such a reverberatory loop to model *short-term* memory storage, which is the type of memory that enables one to dial a telephone number before the number is forgotten. The fact that such a loop can be quickly inhibited does not contradict the properties of this type of memory. The more enduring memories caused by conditioning or learning are, by contrast, modeled by *long-term* memory traces whose sustenance does not depend on an active reverberatory process. Thus, Rashevsky once again faced an issue that is still alive today—namely, whether only one or more than one type of memory exists.

Psychologists contemporary with Rashevsky also faced the dilemma of short-term memory (STM) versus long-term memory (LTM). Hull [115], for example, distinguished between stimulus traces subject to rapid decay and associative strengths able to persist over a longer time period.

Hebb [109] interpreted these memory issues in 1949 with a theory that attempted to bridge psychology and neurophysiology. Concerning reverberatory traces, he said: "Such a trace would be unstable. A reverberatory activity would be subject to the development of refractory states in the cells of the circuit in which it occurs, and external events could readily interrupt

it" [109, p. 61]. Thus, Hebb recognized that to achieve a stable long-term memory, some structural change is needed. He also recognized that the transient trace could be a mechanism for short-term memory: "A reverberatory trace might cooperate with the structural change and *carry the memory until the growth change is made*" [109, p. 62; author's italics].

As a postulate for long-term memory, Hebb formulated the basic concept of classical conditioning in terms of cell interactions: "When the axon of cell *A* is near enough to excite a cell *B* and repeatedly or persistently takes part in firing it, some growth process or metabolic change takes place in one or both cells such that *A*'s efficiency, as one of the cells firing *B*, is increased" [109, p. 64]. As for the nature of the structural change, he proposed that if one cell repeatedly assists in firing another, the knobs of the *synapse*, or junction, between the cells could grow so as to increase the area of contact (see Figure 3). Hebb's associative rule is compatible with a variety of other cellular

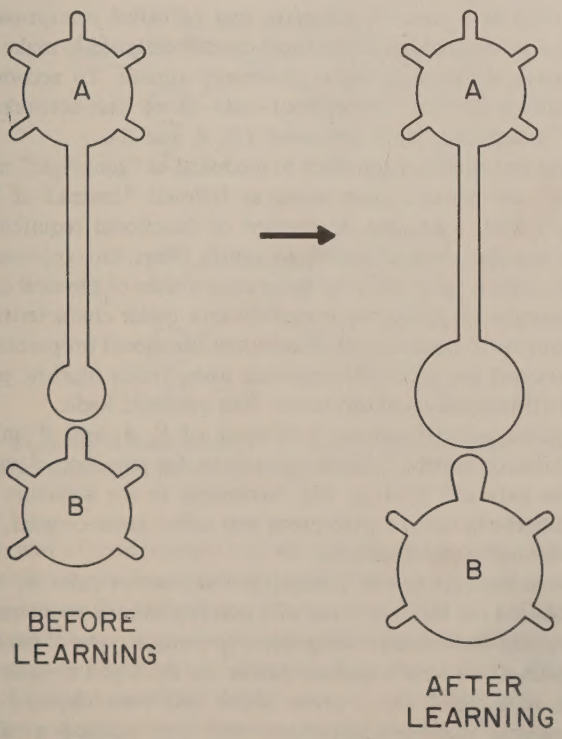

FIG. 3. Diagram of Hebb's hypothesis for structural change. Synaptic knob between *presynaptic* cell *A* and *postsynaptic* cell *B* increases in size after firing of *A* repeatedly followed by firing of *B*.

mechanisms as well. Models built on such mechanisms will be discussed in Section 6.

The next section will examine directions in neural modeling that arose from the digital approach. Later chapters will treat random-net models, and continuous models that incorporate many known features of actual neurons.

3. OFFSHOOTS OF THE DIGITAL APPROACH

3.1. *THE DEVELOPMENT OF THE PERCEPTRON*

The McCulloch-Pitts form of the linear threshold law was generalized to laws whereby activities of all pathways impinging on a neuron were computed, and the neuron fired whenever some weighted sum of those activities was above a given amount. Such laws have frequently been used in neural networks.

In 1962, Rosenblatt [184] synthesized linear thresholds and conditioning rules (see [109]) in a class of automata that he called *perceptrons*. A perceptron was a network of linear threshold devices designed to make and learn choices between different patterns of sensory stimuli. To accomplish this goal, Rosenblatt divided perceptrons into three sub-networks labeled "sensory," "associative," and "response" (S, A, and R).

Rosenblatt described his approach to modeling as "genotypic" rather than "monotypic." He defined these terms as follows: "Instead of beginning ('monotypic') with a detailed description of functional requirements and designing a specific physical system to satisfy them, this approach ('genotypic') begins with a set of rules for generating a class of physical conditions, and then attempts to analyze their performance under characteristic experimental conditions to determine their common functional properties" [184, p. 22]. He developed the genotypic approach using vector algebra, probability theory, and differential equations rather than symbolic logic.

A perceptron was defined as "a network of S, A, and R units with a variable interaction matrix V which depends on the sequence of past activity states of the network" [184, p. 83]. According to the structure of active connections between units, a perceptron was called *series-coupled*, *cross-coupled*, or *back-coupled* (see Figure 4).

A *reinforcement system* was defined as "any set of rules by which the interaction matrix (or memory state) of a perceptron may be altered through time" [184, p. 88]. Rosenblatt distinguished two major types of reinforcement systems: *alpha-system* versus *gamma-system*. In the alpha system, all active connections terminating on a given active cell were changed by equal amounts, whereas inactive connections were not changed at all. In the gamma system, the total value of connection strengths was conserved, so that inactive connections were decreased while active ones were increased.

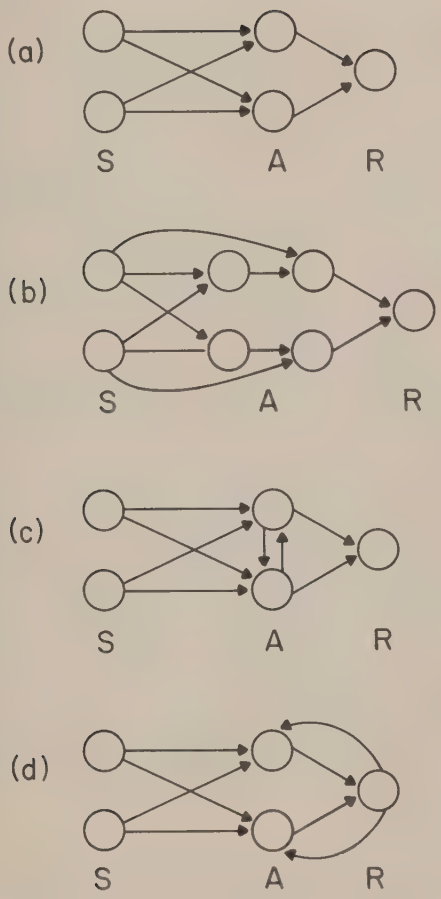

FIG. 4. Examples of some classes of perceptrons: (a) three-layer series-coupled; (b) multilayer series-coupled; (c) cross-coupled; (d) back-coupled.

The amount of reinforcement, η, was determined by one of three *training procedures*. In a *response-controlled* system, the magnitude of η was constant and its sign was determined by the response (that is, by the vector of R-element activities). In a *stimulus-controlled* system, the magnitude of η was again constant but its sign was determined by the stimulus (that is, by the vector of S-element activities). In an *error-correcting* system, η was 0 unless the response was determined elsewhere to be "incorrect."

The first, and simplest, subclass of perceptrons studied by Rosenblatt was that of three-layer series-coupled perceptrons. He discovered that such au-

tomata could be taught to associate an arbitrary coded output, or sequence of outputs, to each stimulus or stimulus sequence in a finite environment. Yet, in spite of their universality, series-coupled perceptrons had major limitations. First, the system had little ability to generalize its response from one stimulus to related stimuli. Thus, it had to learn each stimulus-response association separately, which would take a very long time. Second, it could not separate out parts (*features*) of a complex sensory pattern, which meant that a very large number of coding cells was required. Third, it relied on a reinforcement system controlled by an operator external to the perceptron.

Rosenblatt found that poor generalization and dependence on an external reinforcer could largely be eliminated by adding more connections to the perceptron. This was accomplished either by interposing extra layers of associative units or by cross-coupling existing associative units. Simulation indicated that the system's capacity for detecting features might be improved by back-coupling.

While the specific results of Rosenblatt are not in current use among neural modelers, the issues he raised remain important to the interactions between perception and learning. In particular, there is much current theoretical work on both the organization of sensory feature detectors (see Chapters 7 and 8 of this article) and the design of reinforcement rules (see Chapters 6 and 8). The relevance of these issues is not restricted to models that use perceptron-like error-correcting rules (for example, [3, 7]) or other reinforcement rules in linear threshold devices (for example, [124, 199]), but is basic to all serious multilevel neural population models.

The perceptron also inspired some work in abstract machine theory that was not directly connected to neural modeling. The split between these fields will now be discussed.

3.2. *THE DIVERGENCE OF AUTOMATA THEORY AND NEURAL MODELING*

A sense of how automata theory diverged from neural modeling, while preserving much of the same language, can be obtained from Minsky and Papert's 1969 book, *Perceptrons* [149]. The subtitle, "An Introduction to Computational Geometry," presages the abstract nature of the book. Its approach is to prove general theorems about the recognition and learning of geometric patterns by a particular kind of linear threshold network.

The "perceptrons" of Minsky and Papert were not exactly a subclass of Rosenblatt's. Indeed, the authors stated: "The machines we will study are abstract versions of a class of devices known under various names; we have agreed to use the name 'perceptron' in recognition of the pioneer work of Frank Rosenblatt" [149, p. 4]. Minsky and Papert's machines did not have formally distinct "sensory," "associative," and "response" areas, although they contained parts that corresponded loosely to those loci. The machine

started with a *retina*, which was a grid consisting of small squares, each of which was at any time active or inactive ("light" or "dark"). Downstream from the retina were cells that computed *partial predicates*. Each partial predicate output a value of 1 or 0 based on some rule depending on the activity or nonactivity of the cells in a given subset of the retina. The maximum size of that subset over all predicates was called the *order* of the perceptron. Finally, there was a decision-making cell that computed a linear function of those predicate outputs and responded when that linear function was above some threshold.

The theorems in the early and middle parts of the book dealt with how perceptrons can be chosen to respond selectively to activity patterns on the retina with a given property. When possible, the order of the perceptron was chosen to be bounded as the retina size got large. Examples of such properties were square (pattern *is* a square), square-in-context (pattern *contains* a square), convex, and connected. Toward the end of the book, it was shown that given any partition of all possible binary patterns on the perceptron's retina into two classes, an algorithm existed by which a perceptron could learn to respond to one class of patterns and not to the other. This algorithm worked by changing the coefficients of the linear threshold functions when the machine made an incorrect response. Minsky and Papert did not claim that their learning mechanism was a model of neural learning.

Minsky and Papert's book was a prototype of an approach to artificial intelligence that is often called *sufficiency analysis* [161]. This approach suggests that if a program can satisfy certain heuristic requirements, then the program probably has biological relevance. On this basis, many artificial-intelligence workers believe that heuristic programs for a variety of tasks—for example, geometric-pattern classification, language understanding, or checkers-playing—can be viewed as contributions to psychological theory. By contrast, the neural-modeling literature suggests that different mechanistic realizations of similar heuristics can predict and explain significantly different bodies of data. From this perspective, a sufficiency analysis seems insufficient to judge a model because it provides no criterion of what data properties the model should be able to explain.

As neural modeling became separate from abstract machine theory, it also developed mathematical formalisms quite different from those used by machine theorists. Arbib, who has done much research in both fields, said: "I believe that little of the subject matter of current conferences on automata theory or theory of computation—even such classic concepts as Turing machines, regular events, or context-free languages—will prove applicable in neural modeling" [25, p. 282]. He went on to exhort neural modelers to "return to the *spirit* of the early automata theorists—rather than to apply the *theorems* of current automata theory" [25, p. 291]. Indeed, it will be seen in Sections 6, 7, and 8 that the network-design questions that vexed the

cybernetic pioneers have been partially answered by models based on differential or difference equations.

In designing machines to perform specific classes of tasks, artificial-intelligence workers may still have much to learn from nature. Grossberg (personal communication) has argued, for example, that the current approach of building separate machines for a variety of different visual pattern-recognition tasks is less efficient than building a single machine along the lines of a unified neural theory of vision. This is still a disputed issue among computer-vision specialists (cf. [30, 143]). In general, the question of whether the different goals of neural modeling and artificial intelligence can be achieved by the same theories remains an open one.

The next section will deal with branches of neural-network modeling that grew from the early continuous approaches. In particular, it will discuss how randomness is averaged out in large aggregates of neurons.

4. OFFSHOOTS OF THE CONTINUOUS APPROACH

4.1. *RANDOM-NET MODELS*

The pioneering work on continuous models discussed in Section 2.2 inspired a generation of models that incorporated known neural phenomena into large networks of neurons connected more or less at random. One of these phenomena was the refractory period, in which a neuron that has just had an action potential must remain inactive. Another was the graded (not all-or-none) potentials that occur at the dendrites of a neuron in response to all-or-none action potentials in other cells connected to it. (A historical outline of some of the relevant experimental findings appears in [121].) By Rosenblatt's definitions (see Section 3), these random networks tended to be genotypic rather than monotypic.

Many of the early random-net models were discussed in the last sections of the 1968 review article by Harmon and Lewis [106]. Rich aggregate properties were possible even in networks with only excitatory connections and no inhibitory ones. For example, Rapoport [178] showed that a network with only excitatory connections could "ignite" (become active) depending on two parameters denoting neuron threshold and connection density. Beurle [35], in a continuous model, found that such a network could propagate plane or spherical waves of activity, and that threshold changes in the network could cause shifts in wave peaks analogous to attention shifts in the organism.

The historical process by which inhibition was incorporated into random networks was traced in [106, pp. 576–577] and is summarized here. All-excitatory networks, which were unrealistic from the viewpoint of known neuroanatomy, also exhibited unrealistic patterns of electrical activity. As times became large, the excitatory nets developed both by Beurle [35] and by Ashby et al. [29] tended to approach one of two extremes of activity:

maximal activity leading to saturation of the entire net, or quiescence. The intermediate level of activity found in actual brains was not modeled by these nets. In later articles [36, 37], Beurle conjectured that stable submaximal activity would be possible if inhibition were included. That conjecture was proved in 1963–1965 by Griffith [72–74].

In the years following Griffith's articles, other modelers tried to develop general theories for random neural networks with both excitatory and inhibitory connections. Most of these theories were based on differential or difference equations combined with probability. Some theories borrowed specific formalisms from other scientific fields, with mixed success. Cowan [49] developed a Hamiltonian for a neural network that obeyed the laws of statistical mechanics. Hamiltonians, however, represent a quantity that is being conserved, such as energy, and biologists came to believe that neural nets are dissipative rather than conservative. Consequently, other neural-network theories appeared that were inspired by the dissipative structures of nonequilibrium thermodynamics [64, 120, 175]. Catastrophe theory [201, 219] has been suggested as an alternative formalism, since it can account for hysteretic and discontinuous effects such as those occurring in neurobiology. The framework of the "seven basic catastrophes," however, depends on assumptions likely to be invalid for many neural networks. One of those assumptions is that the system of equations describing the network is a gradient system. Moreover, hysteresis can occur in a wide variety of dynamical systems and is not dependent on catastrophes.

Further evolution of random nets was inspired less by specific mathematical structures than by neuroanatomical and neurophysiological data. In particular, some data indicated that brain connections are random *within* certain neural populations and specific *between* these populations.

4.2. *MODELS THAT COMBINE RANDOMNESS AND SPECIFICITY*

As was mentioned in Section 2, the classical experiments of Lashley [134] showed that many psychological functions, such as specific memories, are retained after extensive brain lesions. Other experiments indicated that specific connections are important for other functions. Mountcastle [154], for example, found that the sematosensory (touch-sensitive) area of the cerebral cortex has a well-organized topographic encoding of the body. Hubel and Wiesel [113, 114] similarly found that cells in the visual area of the cortex are organized into columns that code specific retinal positions or line orientations.

Many modelers have reconciled these two sets of data by means of a principle described as "randomness in the small and structure in the large" [21]. This section will consider models whose equations are explicitly based on this principle. The same principle is implicit in other models to be

discussed in later chapters [7–9, 56, 90, 92–95, 97, 103, 214, 215]. The latter models use purely deterministic equations at the population level that reflect the averaging over large ensembles of stochastic effects at the single-cell level.

The article of Anninos et al. [21] is one of a series of articles including, for example, Harth et al. [107], Anninos [22, 23], and Wong and Harth [216]. In the models described here, the neurons were organized with random connectivities into "netlets," and the netlets, in turn, organized deterministically into larger nets. Evidence for such a netlet was found, for example, in the columnar organization of the visual or somatosensory cortex [113, 114, 154].

Using many cell properties such as refractory periods, Anninos et al. [21] derived an expression for the expected *activity* (defined as fractional number of neurons firing) $\langle \alpha_{n+1} \rangle$ of a netlet at (discrete) time $n+1$ as a function of activity α_n at time n. The crucial variable for determining long-term behavior was a parameter η describing the number of excitatory postsynaptic potentials (here from within a netlet) needed to cause a cell to fire in the absence of inhibitory inputs. If η was very small, $\langle \alpha_{n+1} \rangle > \alpha_n$ for α_n below a certain critical value and then $\langle \alpha_{n+1} \rangle < \alpha_n$ for α_n above that value; thus netlet activity tended to a unique stable steady state. If η was very large, $\langle \alpha_{n+1} \rangle < \alpha_n$ for all values; thus, netlet activity always tended to 0. There was a middle range of η for which $\langle \alpha_{n+1} \rangle < \alpha_n$ for α_n, $\langle \alpha_{n+1} \rangle > \alpha_n$ for intermediate α_n, and $\langle \alpha_{n+1} \rangle < \alpha_n$ again for large α_n. In the latter case there were two stable steady states, one quiescent and one active, and a threshold existed for reaching the active state.

Anninos [22] pursued these principles of network organization further with simulations of multinetlet nets. He found, for example, that a network with random connections could exhibit hysteresis cycles in the dependence of activity of a single netlet on the amount of some external input. He hinted that such hysteresis could be a mechanism for short-term memory, but did not discuss a memory mechanism in detail.

Amari [4–6] described random networks by differential or difference equations with two variable parameters—averaged connection weight and averaged threshold. Depending on the values of these two parameters, the network could have either a single stable steady state, many, or none. If the system had excitatory and inhibitory subnetworks, there could be oscillations of very long period. Association of ideas was also modeled in Amari's system by means of the connection weights.

Freeman [62] laid out some general principles for forming waves from pulses in large neural masses. The differential equations leading to these phenomena were those of a linear feedback system with state-dependent gain [61]. This approach led to models of the olfactory system, in particular to models of positive feedback loops within it. Other models included a hierarchy of levels of organization, from the single neuron through the all-excitatory or

all-inhibitory population to the mixed aggregate and finally to the whole brain.

In later work [63, 64], Freeman discussed how this neural-mass theory could be used to model EEG (brain wave) patterns and predict their frequencies. He described possible techniques:

> Among the more promising approaches is the application of concepts and techniques... of irreversible thermodynamics... in which it is assumed that locally dense synaptic interactions among large numbers of neurons give rise to co-operative activity over neural ensembles. The ensemble activity is... conceived to exist in the form of time varying spatial patterns called neural "dynamic patterns" or "dissipative structures."... The dynamics that underlie these patterns can be described with nonlinear partial differential equations. [64, pp. 348–349].

Work on random models has continued to the present (for example, [38, 190]). However, since about 1975, the work of many active neural modelers (including random-net modelers like Amari and Cowan) has turned from stochastic neural nets to (apparently) deterministic nets. In other words, they have changed their focus from "randomness in the small" to "structure in the large." They have followed Rashevsky by averaging stochastic effects at the single-neuron level into nonlinear, often sigmoid, input-output functions at the neuron-population level. A mathematical justification for this averaging process was given by Geman [67, 68] using stochastic differential equations. Thus, while contemporary neural modelers agree that stochastic effects exist in the brain, the equations to be considered in the remaining chapters are largely deterministic.

In deterministic networks, the specific connections between neuron populations and the rules for interpopulation interactions both play a major role in determining network behavior. The next chapter will examine networks whose connections were suggested by known anatomy of particular brain areas. Specific rules for interactions between populations will be considered in later chapters.

5. MODELS OF STRUCTURE AND FUNCTION IN PARTICULAR BRAIN AREAS

By the 1960s, considerable information existed on the varieties and interconnections of cell types in many regions of the brain. Much knowledge had also been obtained that implicated some of these areas in specific psychological functions. For example, the midbrain reticular formation was known to be involved in nonspecific arousal, the cerebellum in motor control, the older part of the cerebral cortex (archicortex) in emotional expression, and the newer part of the cortex (neocortex) in complex sensory processing and in the generation of goal-directed act sequences. The brain and spinal

cord are so richly interconnected that such behavioral functions are actually mediated by pathways rather than by isolated "centers." Nonetheless, the anatomical and physiological data are sufficient to suggest which anatomical regions lie along which behavioral pathways.

The aim of much neural modeling has been "to hasten the emergence of a coherent set of theories of the regions of the brain and their interactions" [25, p. 291]. Network analyses of specific brain regions have been approached in two different ways. One approach starts with the detailed anatomy of a particular brain area and attempts to correlate the known structure with one or more specific behavioral functions. The other approach starts with psychologically based principles of neural design and derives minimal networks to satisfy those principles. The networks thus derived are then shown to contain anatomical and physiological analogs of specific brain areas. Examples of both approaches will be given and compared in this section. Models of *interactions* between brain regions will be considered in Section 8.

5.1. *THE CEREBELLUM: PATTERN DISCRIMINATION AND MOTOR CONTROL*

The cerebellum has monosynaptic pathways to and from the spinal cord. Also, people or animals with cerebellar damage often exhibit jerky or uncoordinated movements. These facts have led to models relating cerebellar anatomy and physiology to the control of both motor actions and posture-maintaining reflexes.

The anatomy of the cerebellum, which is better understood than that of most other brain areas, is summarized in Figure 5. The Purkinje cells are the largest cells in the cortex (outer layer) of the cerebellum. Their output, which is inhibitory, is the only connection from the cortex to the deeper cerebellar nuclei. Two sets of nerve fibers project to the Purkinje cells: the climbing fibers and the parallel fibers. The climbing fibers (CF) project in a one-to-one fashion from the inferior olive, a midbrain structure that receives inputs from both the spinal cord and the cerebral neocortex. The olive also sends collateral fibers directly to the deep cerebellar nuclei. Parallel fibers project in a many-to-one fashion from cerebellar granule cells, which, in turn, receive the mossy fibers (MF) from other brain areas.

The physiology of these fibers was discussed by Eccles: "The perplexing feature is that the MF and CF pathways carry to the cerebellar cortex much the same information from the periphery or from the cerebral cortex" [54, p. 333]. Most explanations for this dual-pathway system have proposed that one pathway can modify the responsiveness of the other pathway, either in the short term or in the long term.

Marr [140] modeled the cerebellar cortex as a linear threshold network that learned to perform sequences of motor acts in the appropriate stimulus contexts. He said: "the purpose of the cerebellum is to learn motor skills, so

that when they have been learned, a simple or incomplete message from the cerebrum will suffice to provoke their execution" [140, p. 438].

In the model of [140], the unlearned part of the cerebellar apparatus lay in the olivary–climbing-fiber-Purkinje-cell system. A given motor action was treated as a composite of "elemental movements," and each olivary cell then responded to a cerebral instruction for one elemental movement. The one-to-one relation between olivary and Purkinje cells via the climbing fibers communicated these instructions to the cerebellum. Learning took place here at modifiable synapses between parallel fibers and Purkinje cells. (The

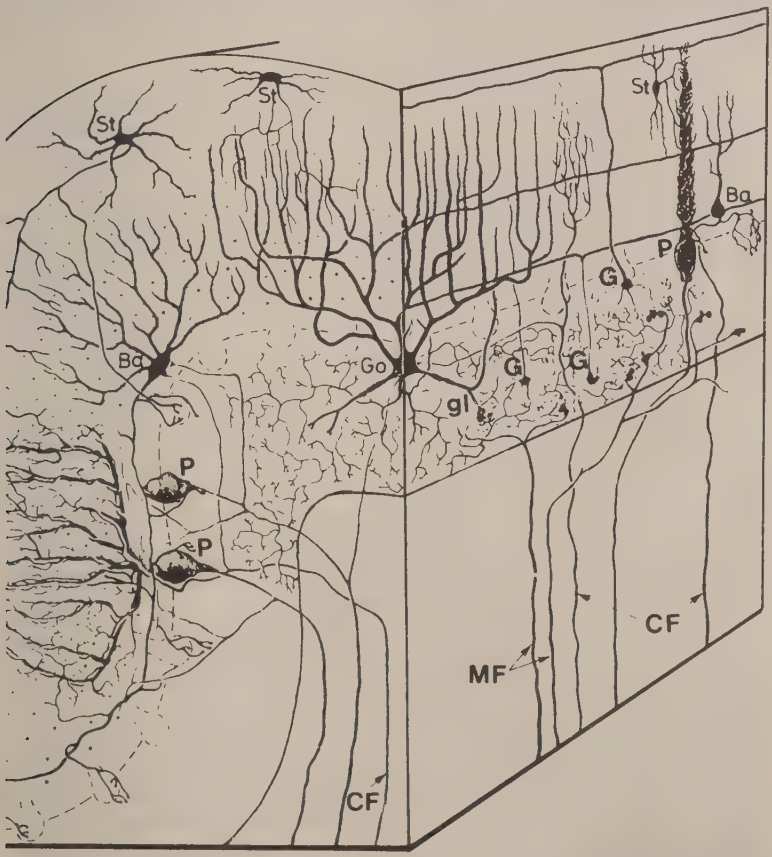

FIG. 5. Schematic of cerebellar cortical anatomy. P = Purkinje cells, MF = mossy fibers, CF = climbing fibers, G = granule cells, gl = glomeruli (where MF synapse with G), and there are three types of small inhibitory interneurons (St = stellate, Ba = basket, Go = Golgi cells). (Modified from Eccles et al. [55]; printed with permission from Springer–Verlag.)

mathematics of modifiable synapses will be discussed in Section 6.) The mossy fiber-granule cell system discriminated sensory patterns and so defined sensory contexts in which movements should be evoked. The mechanism for pattern discrimination was that each granule cell responded to activity in a given subset of mossy fibers called a *codon*. (The word may have been inspired by analogy with the genetic code.)

Marr's model was all-or-none at the cellular level and relied on various counting arguments about numbers of connections. Thus, it omitted detailed description of firing patterns. Also, it did not include a description of how the cerebellum learned any specific movement sequences. Computer simulations of specific motor sequences in a linear threshold cerebellar model similar to Marr's were performed by Albus [2, 3].

A different cerebellar model, emphasizing pattern discrimination rather than motor learning, was developed by Grossberg [87]. In this model the analogy with cerebellar anatomy arose from an effort to build a network that would discriminate a variable pattern. In Grossberg's model, as in Marr's, the only modifiable synapses were between parallel fibers and Purkinje cells. His functional explanation of this fact was subtly different from Marr's: "The mossy fiber input biases the nuclear cells to fire in response to prescribed patterns in the climbing fiber channel" [87, p. 56].

Grossberg hypothesized that the Purkinje-cell input to cerebellar nuclear cells provides a template pattern that can be learned via the parallel-Purkinje synapses. Since the Purkinje output is inhibitory, excitatory inputs from the olivary cells to the nuclear cells are compared with the template via subtractive or shunting inhibition. (The model depends crucially on the olivo-nuclear projections, which are known to exist [55] but were not mentioned by Marr.) Thus, in [87], a collection of nuclear cells emitted a signal if a stimulus pattern was detected that the Purkinje cells had previously learned (i.e., that was "expected"). The psychological implications of expected versus unexpected events will be discussed in Section 8.

Boylls [43] developed an alternative cerebellar model based on cellular data showing that CF electrical activity is immediately followed by a temporary *inactivation*. Such inactivation of specific climbing fibers removes Purkinje-cell inhibition from specific cerebellar nuclear cells. In this model, lateral interactions between nuclear cells corresponding to different locations (cf. Section 7) caused the activity pattern at the nuclei to last longer than the CF inhibition. Since the nuclei contain a topographic encoding of the body muscles to which they project via the spinal cord, this disinhibition encoded the learning of motor patterns. The Boylls model, thus, differed from those of Marr, Albus, and Grossberg in that the cellular changes involved short-term ($\approx$150 millisecond) alterations in cell firing patterns rather than permanent alterations at synapses.

Pellionisz and Llinas [168, 169] proposed that the cerebellum acts as a tensor transforming coordinates in one vector space to coordinates in another space (for example, from head-position coordinates to eye-movement coordinates). They did not give an experimental justification for the assumed linearity of cerebellar transformations. An extension of this model [170] dealt with changes over time in such transformations, without including alterations at synapses. Synaptic change, however, seems to be largely confirmed by recent physiological data summarized by Eccles [54] and Ito [116]. Many of these data were on cerebellar mediation of the *vestibulo-ocular reflex*, that is, of eye movement in response to head rotation. (A detailed discussion of this reflex appears in [183].) Further experimental results should enable comparison *between* different models which include synaptic change.

5.2. *THE RETICULAR FORMATION AND THE LIMBIC SYSTEM: MODE SELECTION, DRIVES, AND COGNITIVE MAPS*

The anatomy and physiology of the midbrain reticular formation has been largely unchanged from reptiles through lower mammals to humans. This suggests its importance in basic survival mechanisms. The theoretical article of Kilmer et al. [122] proposed that the reticular formation (RF) selects the gross behavioral mode for the organism (eating, drinking, sleeping, sex, and so on). This hypothesis was based on a variety of experimental evidence. For example, the RF is involved in initiating the waking or aroused state of the organism [153], and RF lesions lead either to disintegrated behavior or pathological sudden mode switches. Also, the RF is the only large area of the central nervous system that is monosynaptically connected to every other large area.

Kilmer et al. did a computer simulation of the reticular formation via a network called S-RETIC. S-RETIC consisted of independent yet connected modules, analogous to the 100-micron-thick "poker chips" into which the actual RF is divided [186]. Each module was a nonlinear probabilistic computer which constantly received information from input indicators and from other modules (cf. Figure 6), and continually recomputed the probability of assignment of the organism to a given mode. Since different modules received different sensory information, their mode probabilities at any given time were unequal. The organism reached a decision when more than half of the modules were assigned to the same mode, each with a probability of at least one-half.

The multiple-module organization had two advantages. First, the organism could receive sensory information from a broad area. Second, the von Neumann principle of reduced vulnerability to malfunction was upheld.

The S-RETIC used by Kilmer et al. decided on modes but did not commit the organism to specific actions within those modes. On more sketchy

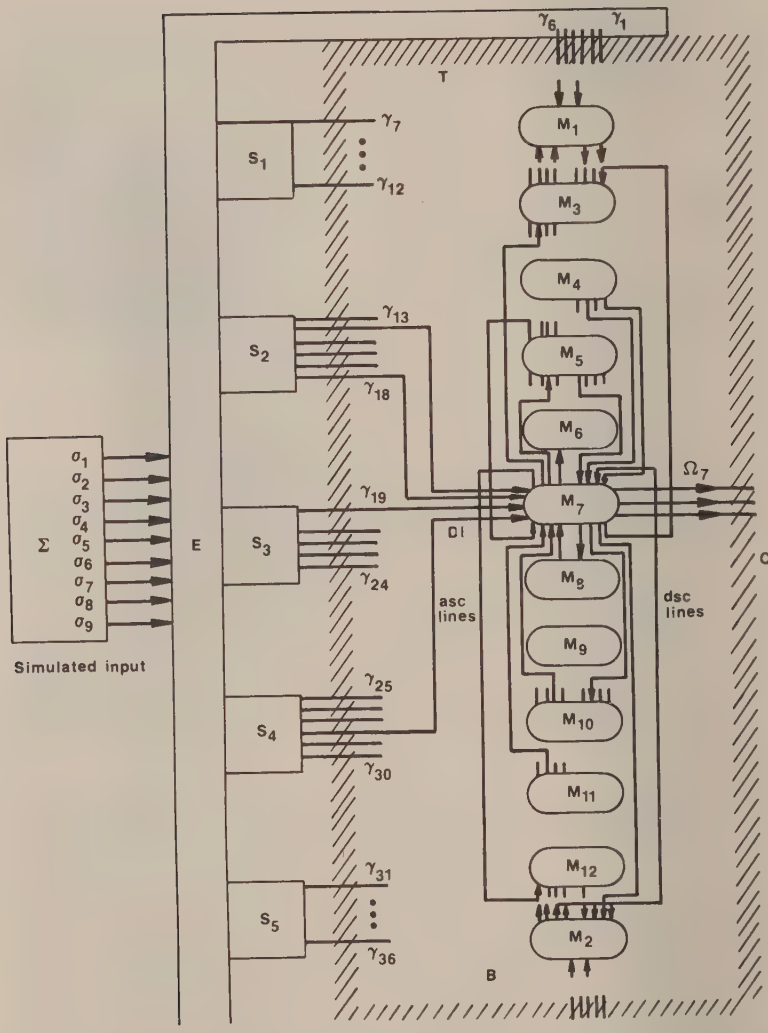

FIG. 6. S-RETIC simulation model: S_i were input systems, M_i were modules, Ω_i were S-RETIC's mode-indicating output. Connections are shown only to one module M_7. Lines "asc" and "dsc" meant ascending and descending lines. σ_i and γ_i were binary. (From Kilmer et al. [122]; reprinted with permission from Academic Press Inc. (London) Ltd.)

evidence, decision within a mode was assigned to the hippocampus (part of the archicortex) by Kilmer and McLardy [123] and Kilmer and Olinski [124]. Decision-making was mediated by inhibitory hippocampal outputs corresponding to competition between possible acts. This competition was very similar in organization to the competition between modes in the simulated RF.

Whereas S-RETIC simulated instinctive behavior, the Kilmer-McLardy and Kilmer-Olinski hippocampal models had much larger learned components. Linear threshold networks with modifiable connection weights were employed to simulate the process by which pyramidal cells (the largest cells present) learn firing patterns in response to sensory data. Learning consisted of positive or negative reinforcement of a given act based on whether that act satisfied or frustrated the drive relevant to the mode in operation. Reinforcement was mediated by inputs from an area corresponding to the septum (part of the brain's subcortical limbic system).

The notion of mode selection discussed by Kilmer et al. had much in common with that of Grossberg [91]. Grossberg proposed a mode selection mechanism based on a *combination* of strength of internal drive, nonspecific arousal, and sensory inputs. After the organism had chosen which drive to satisfy, it could decide among motor actions on the basis of positive and negative reinforcement.

Grossberg's approach differed from Kilmer's in that "the network regions are not presumed to be exact replicas of real anatomical fragments. Nonetheless, the anatomical relationships between the network regions, as well as their functional roles in total network processing, suggest natural analogs with real anatomies" [91, p. 272]. For example, the drive and arousal loci were qualitatively analogous to the hypothalamus and reticular formation, and the sensory area to the mock-neocortex. Another locus, which supplied motivational feedback to the sensory area, was analogous to the hippocampus; this mock-hippocampus fed and was fed by still another locus implicated in reinforcement (mock-septum). This model will be discussed more fully in Section 8.1.

Marr [142] posited a temporary memory storage role for the hippocampus. His model used all-or-none neurons and was structurally similar in organization to his previous models of the cerebellum [140] (discussed in Section 5.1) and the neocortex [141] (to be discussed in Section 5.3). He proposed that the hippocampus was only capable of storing memories, and not classifying them like the neocortex.

The model of [142] used codons like those of [141] for memory storage. An input pattern arrived at one group of cells, and there were modifiable synapses in both directions between the input cells and the codon-forming cells. The modifiable synapses used were of types developed by Brindley [45]

and Hebb [109], which will be discussed in Chapter 6. Some histological evidence was adduced that input cells could be hippocampal pyramidal cells, and codon cells could be in the entorhinal area of the hippocampus.

Stanley (see [27, p. 126]) found from computer simulation that Marr's hippocampal model would need extremely large numbers of cells to be effective. Also, recent data summarized by O'Keefe and Nadel [162] indicate that the hippocampus may not be involved in all memory, but only selected kinds of memory. These authors fit a variety of behavioral studies on hippocampally lesioned nonhuman mammals and hippocampally damaged human patients into an overall theory. This theory was that the hippocampus is important in the formation of a cognitive map of physical space, as in coding the location of rewarding or punishing stimuli. They did not suggest a network for implementing such a map; the integration of their theory with existing network models will be considered in Section 8.1.

5.3. *THE CEREBRAL NEOCORTEX: PATTERN CLASSIFICATION AND GOAL DIRECTION*

The cerebral cortex (neocortex) is the outermost and evolutionarily youngest part of the brain. The neocortex of humans and other primates consists of six layers with rich interconnections. That of lower mammals consists of six layers in some parts and two layers in others, whereas reptiles have only a rudimentary neocortex. Greater cortical complexity reflects an increased ability to make associations between sensory stimuli and to initiate complex goal-directed behavioral sequences.

The neocortical model of Marr [141] contained many of the same kinds of components found in Marr's cerebellar model—large output neurons (here the pyramidal cells), small inhibitory interneurons, granule cells, and so on. Again, these were used to create codons. However, the codons were not all-or-none as in the cerebellar model. In this model, the cerebellum could only recognize the presence or absence of a particular sensory pattern. In contrast, the neocortex could generalize from one pattern to related ones. This increased complexity was achieved by making additional synapses modifiable. The variable synaptic efficacies served as probabilities for certain associations. The types of modifiable synapses used were those of Brindley [44, 45] (see Section 6.1). Whereas the cerebellum, in Marr's model, contained only one type of codon recognizer (the granule cells), his cerebral cortex contained two (the granule and Martinotti cells). Synapses from both types of codon recognizers to the large pyramidal cells were modifiable.

The model neocortex of [141] was also more complex then the model hippocampus of [142] in that it could classify patterns as well as store them. This classifying ability was made possible by climbing fibers that enabled one cell to sample many others at a time.

Scott [188, 189] developed a series of models for the neocortex, based on the observation that the structure of the cortex is largely unchanged over its surface. This structural homogeneity seemed paradoxical, since function varies: some parts of the cortex process data from one specific sensory modality, while other parts associate data from different modalities. Nonetheless, the structural unity was interpreted in terms of a certain unity of function, in that all parts of the cortex were involved in active feedback between goals and percepts. Various predictions on cell types (of which more had been found since Marr's article) were made on the basis of simulations of such feedback loops.

These general neocortical models did not make the needed distinctions between sensory and association cortices, distinctions that are based on connection patterns with other cortical or subcortical areas. Sections 7 and 8 will discuss several models of pattern processing in the visual cortex. The association cortex (that is, frontal, temporal, and parietal lobes), is perhaps the hardest area in the brain to model because it integrates data from all the sensory cortices and (in the case of the frontal lobe) from other organs of the body via the hypothalamus. Yet, enough experimental data exist on the behavioral functions of these areas to suggest models; for example, the parietal lobe is implicated in attentiveness and in organizing motor commands [155, 156], and the frontal lobe in forming temporal structures of motivated behavior [66].

5.4. *GENERAL REMARKS ON MODELS OF BRAIN AREAS*

The further one goes in the brain from either the sensory or the motor end of pathways, the more ambiguity there is in assigning structure to function. In such "interneuronal" areas as the reticular formation, limbic system, cerebellum, and association cortex, much has been established about anatomy, but less about cellular physiology. The state of experimental development is reflected in the development of models.

Both structure-based and function-based models of brain areas are in active use among contemporary modelers. Chapter 8 will discuss global brain organization as it relates to such complex multipathway processes as conditioning and code development. The multilevel networks used to model such processes have suggested function-based models of individual levels. Such models have led to detailed experimental predictions, the results of which could suggest refinements of the models that are more structure-based. Thus, understanding "regions of the brain and their interactions" requires feedback between the two types of models.

Before multilevel models can be discussed, it is necessary to have some understanding of typical components of large models. The next two chapters will be devoted to detailed consideration of two such components: modifiable synapses and lateral inhibition.

6. MODIFIABILITY OF SYNAPSES IN LEARNING MODELS

Ever since the work of Hebb [109], the most common neural hypothesis used to explain learning or conditioning has been that some functional change takes place in one or many synapses between neurons (cf. Figure 7). Most theorists (and many experimentalists) believe that at some synapses in the brain, correlated activity at the pre- and postsynaptic cells alters the efficacy of the synapse in causing action potentials at the postsynaptic cell.

The primary evidence for synaptic change (otherwise known as *plasticity* or *modifiability*) comes from psychology rather than neurophysiology. Indeed, apart from data about the mechanisms that regulate the growth of synaptic connections during development [119, 173], few physiological data are available to directly verify the existence of plastic synapses. Rather, the concept emerges from the vast psychological literature on conditioning. Hebb's contribution was to interpret known psychological operations in terms of neural anatomy. This interpretation continues both to stimulate an experimental search for the locus of memory at synapses (see, for instance, [71, 118, 202, 217]) and to provide a unifying construct to aid in explaining other types of neural data.

As discussed in Chapter 2, Hebb [109] proposed that changes in the efficacy of synapses could take place via growth of synaptic knobs. A variety of alternative mechanisms has also been proposed for changes in efficacy. Such mechanisms have included, for example, correlated changes in the amounts of usable presynaptic transmitter substance and of postsynaptic protein synthesis [80]; changes in the conformation of a receptor protein at the postsynaptic membrane [197]; and resistance changes at the postsynaptic membrane [138, 218].

Processes of directed cell growth, enhanced transmitter production, and altered site sensitivity are likely to obey the same formal rules. Hence, before the exact biochemical mechanism is identified in a particular experimental

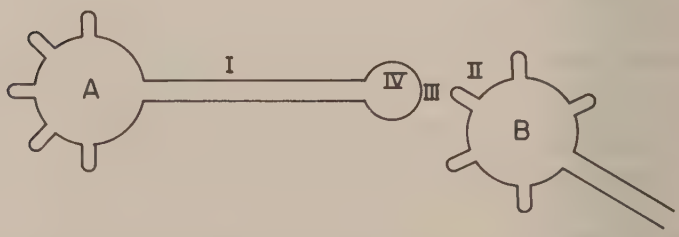

FIG. 7. Schematic drawing of the most common kind of synapse from pre-synaptic neuron *A* to post-synaptic neuron *B*. I: axon of *A*; II: dendrite of *B*; III: synaptic gap; IV: synaptic knob.

situation, it is useful to study the properties of networks containing synapses that are modifiable according to given rules.

6.1. *DISCRETE SYNAPSES WITH ASSOCIATION RULES*

The most common formal rule used by modelers of plastic synapses is an *associative* rule. Such a rule means that presynaptic electrical activity followed after some time delay by postsynaptic electrical activity creates a *tendency* to increase the efficacy of the synapse. Such a rule is sometimes called "Hebbian," but the term "associative" is preferred here because the rule can have subtle consequences not discussed by Hebb [109]. For example, with an associative rule, efficacy of a synapse can sometimes decrease rather than increase when pre- and postsynaptic activities are paired. This can occur if the decay rate of synaptic efficacy is large relative to the amount of pairing, or if the pre- and postsynaptic elements are embedded in a suitable network with mutual inhibition.

Some simple discrete neural networks that mimic aspects of associative learning were proposed by Brindley [44, 45]. These networks used all-or-none model neurons, and linear thresholds, with the addition of modifiable synapses and some random elements. Brindley followed McCulloch and Pitts [147] in basing the mathematical description of his network on symbolic logic.

In [44], ten types of plastic synaptic knobs and the logic of their operation were discussed. Of the ten, the most important was the "Hebb synapse" where facilitation depended on correlated pre- and postsynaptic activities. Also, it was possible for two neurons in the model to have two synapses between them, one modifiable and one unmodifiable.

Figure 8 shows learning networks that contained these types of synapses. These networks modeled aspects of *classical* or *Pavlovian conditioning* (the type that occurs when a dog learns to salivate in response to a bell after the bell has been repeatedly paired with food) and *operant conditioning* (the type that occurs when a rat learns to press a lever after the act of lever pressing has been repeatedly rewarded with food). The all-or-none character of neurons oversimplified the description of conditioning data. The networks used to model such data, however, had structural details in common with certain continuous models. For example, some structural analogies can be drawn between Figure 8(a) and Figure 23. (The latter figure is derived from the continuous model of Grossberg [85] and will be discussed in Section 8.1.) All-or-none elements with one modifiable and one unmodifiable synaptic knob are analogous to continuous neural elements where synaptic efficacy can take on a range of values.

Brindley noted some possible ways of making his discrete model behave more like a continuous model. For example, he stated that for the network to

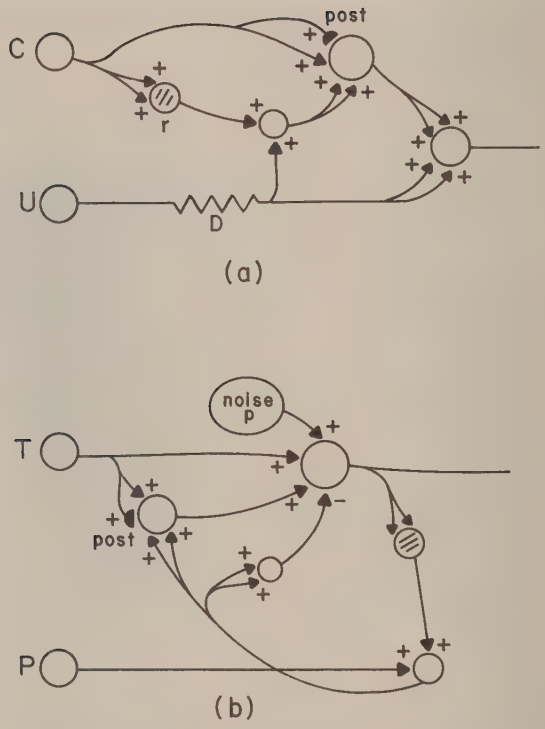

FIG. 8. (a) A network for classical conditioning. Conditioned stimulus (CS) was followed by unconditioned stimulus (US). Cell γ fired repetitively in response to a single impulse, and D was a delay. Arrows denote unmodifiable synaptic knobs, semicircles denote modifiable knobs. "Post" means the synapse was modified by postsynaptic activity. (Modified from Brindley [44].) (b) A network for operant conditioning. T was trigger stimulus. Before conditioning, T had low probability (p) of provoking response. After response had been repeatedly rewarded with reinforcing stimulus P, that probability rose to nearly 1. (Modified from Brindley [44]; printed with permission from The Royal Society (London).)

be able to forget what it has learned, the efficacy of modifiable synapses must decay over time. That is, when a synapse has been conditioned, it reverts to its initial state either in a fixed time or with a fixed probability. Such a decay was not explicitly included in either of Brindley's articles, but is common in the networks of other modelers to be discussed.

Brindley [45] modeled the learning of a sequence of three words. The network used the same elements as that of [44], with the addition that cells were allowed to have different (but still integer-valued) thresholds. He found

it was easy to generalize the net to one that could learn nonsequential inputs composed of familiar sounds, but much harder to generalize to a net that could learn *combinations* of words.

Brindley's ideas of synaptic modifiability were used by others (for example, Marr [141, 142]) to model more complex data within an all-or-none framework (cf. Sections 5.2 and 5.3). The limitations of his models were those of the all-or-none framework itself.

Associative synapses were also the basis for learning in the linear model of Anderson [16], which was described using vector algebra and probability theory. Anderson's model was based on his own models of memory storage, retrieval, and recognition [14, 15]. In all these articles, a memory trace was described as a vector $\mathbf{f} = (f_1, f_2, \ldots, f_n)$, each of whose components was the activity of a single element in the network. If the elements were neurons, this activity would be instantaneous firing frequency; if the elements were populations, activity would be average firing frequency. All the memory traces present in those elements were summed into a total storage vector

$$\mathbf{s} = \sum_{k=1}^{K} \mathbf{f}_k .$$

The model was discrete in time, although synaptic efficacy could take on any of a continuum of values.

The problems considered in [14] were:

(1) *Recognition*: If an input pattern (vector of activities) is given, how can one tell whether a trace similar to that pattern is already stored as part of the storage vector?

(2) *Retrieval*: Once recognition has occurred, how can the stored trace be reconstructed?

(3) *Association*: Once retrieval has occurred, how can other stored traces be found which are also similar to the input pattern?

Anderson stated: "It is very easy to store information in such a storage unit. The significant problems are involved in the retrieval of the stored information" [14, p. 114]. Storage is easy because in this model, "information" referred to a formal mathematical description rather than a detailed physiological mechanism, and a vector array is easy to describe. Retrieval involved recovering one term of a vector sum, which was subject to error. The techniques described for solving the recognition and retrieval problems used various kinds of filters designed so as to minimize the probability of error—that is, to maximize a certain signal-to-noise ratio. These filters were described in terms of formal mathematical rules rather than specific cellular mechanisms.

The mathematical theory of these linear filters was discussed further in [15]. The input trace $\mathbf{f}$ was used to construct a *matched filter* whose output was the scalar product of the stored array with the input trace, $V = \mathbf{s}\cdot\mathbf{f}$. If $\mathbf{s}$ was considered to be $\mathbf{f}$ plus a noise component $\mathbf{n}$, the signal-to-noise ratio was defined as $(\mathbf{f}\cdot\mathbf{f})^2/[(\mathbf{n}\cdot\mathbf{f})^2]_{\text{avg}}$, the average being taken over all possible average storage vectors. The signal-to-noise ratio was then close to N/K, with N the number of neural elements and K the number of traces. Thus, increasing the number of elements increased reliability, as in [159], and increasing the number of traces made recognition more difficult. The signal-to-noise ratio achieved was the maximum possible for a linear filter, but could be improved by using a nonlinear filter derived from a spatial cross-correlation vector.

The problem of association was discussed in [14], and its relationship with a theory of synaptic connection weights was described in [16]. In these articles, a model for association was proposed which involved two sets of neurons, α and β. The following assumptions were made for simplicity: (1) both α and β had the same number N of neurons; (2) there was another number M such that M neurons of α projected to every neuron of β, and every neuron of α projected to M neurons of β. If there was assumed to be an input trace $\mathbf{f}$ at α, a trace $\mathbf{g} = (g_1,\ldots,g_n)$ at β, and values a_{ij} for the efficacy of the synapse from the ith element in α to the tth element in β, then $\mathbf{g} = A\mathbf{f}$, where A was the matrix of connection weights a_{ij} (see Figure 9). It was shown that the optimal matrix A (from the standpoint of signal-to-noise ratio) was obtained by

$$a_{ij} = cg_i f_j. \tag{6.1}$$

If there were K input traces $\mathbf{f}_k = (f_{k1}, f_{k2},\ldots,f_{kn})$, $1 \leqslant k \leqslant K$, and K output traces $\mathbf{g}_k = (g_{k1}, g_{k2},\ldots,g_{kn})$, all with equal "power" $P = \mathbf{f}_k\cdot\mathbf{f}_k = \mathbf{g}_k\cdot\mathbf{g}_k$, then (6.1) generalized to

$$a_{ij} = c\sum_{k=1}^{K} g_{ki} f_{kj}, \tag{6.2}$$

and normalization considerations led to a choice of $c = N/MP$.

In this model, synaptic weights were calculated instantaneously from the inputs, as specified by (6.2). The temporal dynamics of the system were not developed, and no decay terms were included. Thus, a trace could only be forgotten if there was interference by the memory of another stimulus whose trace involved the same neural elements. Such interference is not a model for selective attention, because an incoming stimulus important to the organism's survival may affect different neural loci than the stimulus currently being attended to.

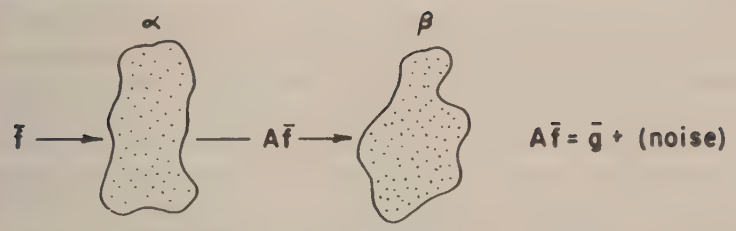

FIG. 9. Scheme for associating distinct memory traces f and g, with α and β two groups of neurons. (From Anderson [16]; reprinted with permission from Elsevier Science Publishing Co., Inc.)

Anderson suggested the inclusion of decay terms in [17]. In this article, he modeled some data on the learning of lists, using the framework of his earlier articles along with some temporal considerations. In the data being modeled [198], a subject first memorized a short list of items; then an item was presented and the subject was asked whether it was on the list. The reaction time was tested and found to be linearly related to the number of items on the list.

The same article also considered other data [200] where the reaction time was a function of the probability of occurrence of an item. Such data were modeled by letting the kth trace coefficient be $c_k = 1 + \gamma_{p_k}$, where γ was a weighting constant and p_k was the probability of presentation of the kth item. Anderson said:

> The probabilistic term due to the immediate past history could easily arise in the following way: Each time the item is tested or rehearsed, the stored trace is slightly strengthened. This can easily occur since the trace has just been present in the correct form. If an exponential decay of trace strength in short-term memory is assumed... the probability dependent term will appear to be of this form. [17, p. 429]

He thus suggested, without proposing a quantitative formula, that a memory trace decays exponentially when the item being remembered is not being presented and is strengthened by presentation.

Nass and Cooper [158] added more explicit neural mechanisms to the linear framework of Anderson's articles. They shows that an optimal matrix of the form (6.1) arose as the limit of a matrix which was modified at time t according to the rule $A(t) = \gamma A(t+1) + \eta(\mathbf{g} \cdot \mathbf{f})$, γ and η positive constants, with the decay rate $\gamma < 1$ but close to 1. This article went on to use the same framework in a model of feature detection in the visual cortex; other aspects of the same issue are discussed in Anderson et al. [20]. In Section 8, those models will be discussed and compared with nonlinear models of the same phenomena.

In more recent work, Anderson has used his type of linear models to simulate a wide range of data in the learning of categories. For example, the network of [18] learned how to classify names into abstract categories (mortal, immortal, man, God, et cetera). The network of [20] (which will be discussed again in Chapter 8) learned distinctions between spoken phonemes where ambiguity was possible. Other forms of learning exist which include nonlinear effects that are hard to explain by a linear model. An example is *serial learning* [165]—that is, learning the *order* of items on a list.

Anderson's theories of associative memory and category learning are similar to theories of the same phenomena by Kohonen and his coworkers [129–131]. The term "associative memory" (or "associative learning") was used in [131] to mean the recollection of a total pattern if part of the pattern is perceived. But the term is also used for associations that can develop via classical conditioning or serial learning. In fact, Kohonen's model, like many others, treats all these aspects of associative learning as part of a unified process.

The mathematical theory of Kohonen's model was described in [131]. There were input spike frequencies ξ_i, $i=1,\ldots,n$; output spike frequencies η_i, $i=1,\ldots,n$; and connectivities μ_{ij} satisfying

$$\begin{aligned} \eta_i &= \beta_i \xi_i + \sum_j \mu_{ij}\eta_j + \eta_b', \\ \frac{d\mu_{ij}}{dt} &= \alpha\eta_i(\eta_j - \eta_b), \end{aligned} \tag{6.3}$$

where η_b' denoted background activity and η_b denoted the average of η_b' over some time period. A network with input and output subnetworks arranged in a rectangular lattice and obeying (6.3) was simulated. These simulations incorporated lateral inhibition (cf. Section 7) between inputs. Such inhibition operated by transforming the input values ξ_i before applying (6.3). Each ξ_i was replaced by a value ξ_i' which was a weighted sum of neighboring input values (for example, $4\xi_i - \sum_{\text{dist}(i,j)=1}\xi_j$).

The simulations involved recognition of human faces by means of distinctive features. The organization of the network suggests that of some of the more complex networks for feature detection [38, 92–94, 139, 158] to be discussed in Section 8.

6.2. *CONTINUOUS SYNAPSES WITH ASSOCIATION RULES*

One of the first modelers (starting in 1967) to translate an associative rule for synaptic modification into explicit equations was Grossberg [75–83]. His formal rules and related equations were derived from psychological pos-

tulates. The networks implementing these rules, in turn, suggested analogs of neural elements and interconnections. The most general form of the equations [81] remains in current use in the modeling of multilevel adaptive networks with modifiable synapses between levels (cf. Section 8).

The psychological and neurophysiological implications of the theory thus derived were described in [77]. Mathematical results about the equations were proved elsewhere (for example, [75, 76, 79, 81–83, 86]). In [77], the following question was asked: How does an organism learn to produce one sound (say B) in response to another (say A) after repeatedly hearing them presented in sequence? The network designed to answer that question incorporated the following psychological principles among others:

(1) Language appears to be spatiotemporally discrete; that is, a sound like A is psychologically treated like an "atom" instead of being subdivided.

(2) Such discrete symbols as occur in language are used to represent sensory experience, which is spatiotemporally continuous.

(3) Learning changes an act from continuous to discrete; for example, a child learning to walk must concentrate continuously on his or her movements, whereas an adult walking has an automatic sequence of discrete steps.

The network for satisfying these postulates consisted of discrete elements with time-varying activities that satisfied continuous differential equations (see Figure 10). For each stimulus "atom" such as A, a variable called the *stimulus trace* $x_A(t)$ was defined that measured how active the memory for A was at any given time t. For each pair of "atoms" A and B, a variable called the *associational strength* $z_{AB}(t)$ was defined that measured how strong the sequential association AB was in the network's memory at time t. Both of these variables were previously used in qualitative learning models [115]. In Figure 10, the ith stimulus trace x_i was located at the vertex v_i and the c_{ij}th association z_{ij} was located along the edge e_{ij}. Grossberg noted that the v_i were analogous to cell bodies, the e_{ij} to nerve axons, and the junctions $e_{ij} \rightarrow v_j$ to synapses.

Different articles by Grossberg used variations on the same general equations for the network of Figure 10. One of the forms of these equations

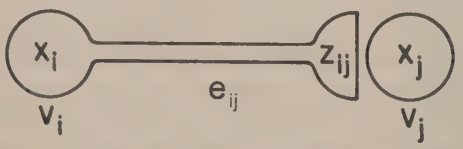

FIG. 10. Schematic of two elements and one modifiable synapse between them in Grossberg model.

used in his later work is

$$\dot{x}_i(t) = -\alpha_i x_i(t) + \sum_{k=1}^{n} \left[x_k(t-\tau_{ki}) - \Gamma_{ki}\right]^+ \beta_{ki} z_{ik}(t) - \sum_{k=1}^{n} \left[x_k(t-\sigma_{ki}) - \Omega_{ki}\right]^+ \gamma_{ki} + I_i(t), \tag{6.4a}$$

$$\dot{z}_{jk}(t) = -\delta_{jk} z_{jk}(t) + \varepsilon_{jk}\left[x_j(t-\tau_{jk}) - \Gamma_{jk}\right]^+ x_k(t), \tag{6.4b}$$

where the I_i were outside inputs, α_i and δ_{jk} were decay rates, Γ_{ki} and Ω_{ki} were thresholds, τ_{ki} and σ_{ki} were time delays, β_{ki} and ε_{ki} were constants denoting signal strengths, and x^+ denoted $\max(x,0)$ for any real x. Thus, the associative strength at any synapse changed with cross-correlated pre- and postsynaptic activities [as seen in the $x_j(t-\tau_{jk})x_k$ terms]. Postsynaptic activity, in turn, changed with the cross-correlation of presynaptic activity with synaptic strength [as seen in the $x_k(t-\tau_{ki})z_{ki}$ terms]. The $-\gamma_{ki}x_k(t-\sigma_{ki})$, which were not present in the earliest versions of these equations, reflect lateral inhibition between vertices (see Section 7).

Some new psychological properties, separate from the original postulates, emerged from the results of (6.4). These were:

(1) The more often the network heard B following A, the more likely it was to say B after hearing A.

(2) An isolated network could remember without practicing overtly.

(3) Memory could sometimes improve spontaneously on recall trials.

(4) If the network had learned the association AB, it could be changed to AC by sufficient presentation of the new sequence, but the change took a long time if the association AB had become very strong.

Grossberg [75, 76] proved theorems about the limiting behavior as t approached infinity of the relative sizes of different x_i and different z_{jk}, and interpreted the limits in terms of long-term memory storage. He made various modifications of (6.4) in succeeding articles. For example, he noted that in (6.4), the association between A and B was weakened while the network was not actively hearing A. This property, which is contrary to intuition, was corrected by replacing (6.4b) with the equation $\dot{z}_{jk}(t) = \{-\delta_{jk}z_{jk}(t) + \varepsilon_{jk}x_k(t)\}[x_j(t-\tau_{jk}) - \Gamma_{jk}]^+$, so that the association strength z_{jk} could only decay τ_{jk} time units after a stimulus trace existed at v_j.

In [75], the limiting behavior of these equations was studied for one type of network geometry called the *outstar* (see Figure 11). In the outstar, one source vertex (or cell, or cell population) v_1 projected to sink vertices $v_2,\ldots,v_n$. Long-term storage was interpreted as residing in the *relative*

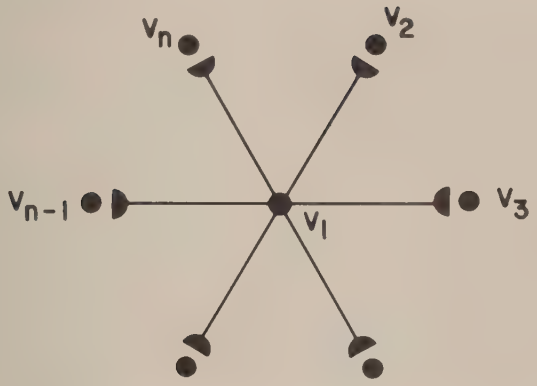

FIG. 11. Outstar anatomy. (Modified from Grossberg [75]; printed with permission from Academic Press, Inc.)

weights of $x_1, x_2, \ldots, x_n$ and of $z_{12}, \ldots, z_{1n}$, i.e., in the functions

$$X_j(t) = x_j(t)\left(\sum_{k=2}^{n} x_k(t)\right)^{-1},$$
$$y_{ij}(t) = z_{ij}(t)\left(\sum_{k=z}^{n} z_{ik}(t)\right)^{-1}. \tag{6.5}$$

In this context, the outstar was the smallest anatomy (that is, network geometry) that could learn a choice; for example, that A was followed by B rather than C, D, E, or Z.

Grossberg showed that for many classes of inputs to (6.4), the functions $y_{ij}(t)$ and $X_j(t)$ as defined by (6.5) were eventually monotone and so approached limits as $t \to \infty$. Moreover, for each j, $\lim_{t\to\infty} y_{ij}(t) = \lim_{t\to\infty} X_j(t)$. Thus, the same distribution of weights was coded both in the relative stimulus traces and in the relative associational strengths at the outstar border. One important case was that in which inputs formed a *spatial pattern*, defined as $I_j(t) = \theta_j I(t)$, $j \neq 1$, $\sum_{j=1}^{n} \theta_j = 1$. In this case the input patterns weights were shown to be stored in long-term memory, that is,

$$\lim_{t\to\infty} X_j(t) = \lim_{t\to\infty} y_{ij}(t) = \theta_j, \tag{6.6}$$

under certain conditions on the inputs $I_1(t)$ and $I(t)$. The meaning of these conditions was that inputs to both the center and the border were presented "often enough" for large times.

If activation of the outstar's center was interpreted as a conditioned stimulus (CS) and a spatial pattern applied to its border as an unconditioned stimulus (US), the learning of a spatial pattern could be seen as an analog of classical conditioning. The limit theorems were extended to some outstars whose borders received inputs only for a finite time period. One example, where only v_1 and v_2 were activated, was analogous to repeated presentations of B after A.

In [76], the investigation of variants of (6.4) was extended to an anatomy called the *complete graph with loops* (see Figure 12). In this anatomy, every vertex projected to every vertex including itself. If inputs were shut off at time $t = 0$, the network could either remember spatial patterns in the same manner as did the outstar, or else forget them. In other words, for some parameter choices, (6.6) held and the differences between X_i, y_{ji}, and θ_i were all eventually monotone (see Figure 13). For other parameter choices, however, the deviation of $\lim_{t\to\infty} y_{jk}(t)$ from $y_{jk}(0)$ could be made arbitrarily small. Thus, if $y_{jk}(0) - y_{ik}(0)$ is large enough, $\lim_{t\to\infty} y_{jk}(t) \neq \lim_{t\to\infty} y_{ik}(t)$, violating (6.5).

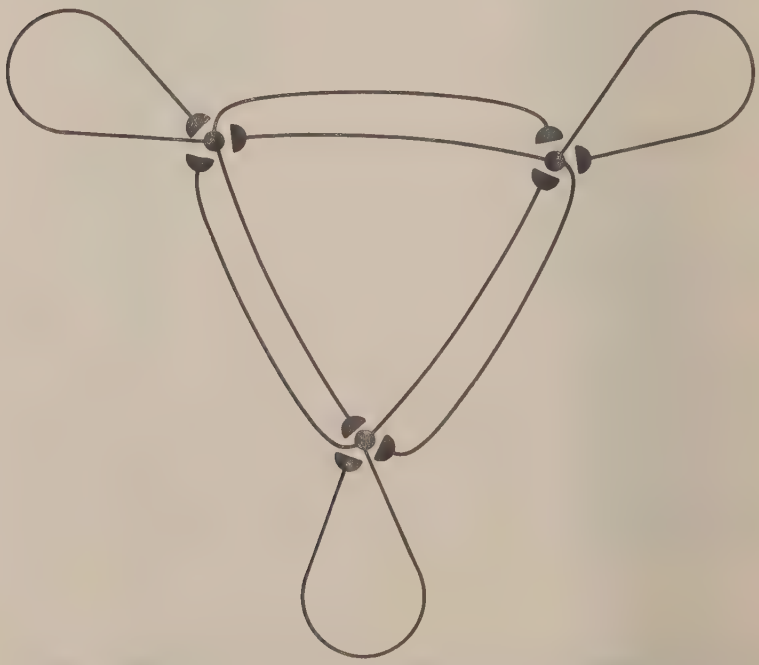

FIG. 12. Complete graph with loops, $n = 3$. (Modified from Grossberg [76]; printed with permission from Academic Press, Inc.)

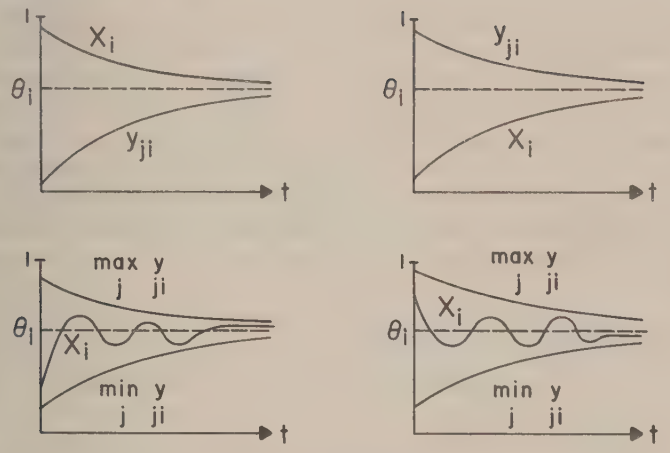

FIG. 13. Possibilities for monotonic approach of y_{ji} to x_i and of both to θ_i. (Modified from Grossberg [82]; printed with permission from Academic Press, Inc.)

This result illustrates that within the same formal mathematical structure, changing one or more network parameters can profoundly alter network behavior. Behavior can also be altered by adding or removing network connections. This does not change the formal structure of the equations, but merely sets some connectivity parameters equal or not equal to zero. For example, Grossberg [82] showed that in a complete graph, *without* loops, pattern weights could be forgotten even for parameter choices that would produce perfect memory if loops were present.

Grossberg [81–83, 86] extended the study of (6.4) and its modifications to the more general network equations

$$\begin{aligned} \dot{x}_i(t) &= A_i(W_t, t)x_i(t) + \sum_k B_{ki}(W_t, t)z_{ki}(t) + C_i(t), \\ \dot{z}_{ji}(t) &= D_{ji}(W_t, t)z_{ji}(t) + E_{ji}(W_t, t)x_i(t), \end{aligned} \tag{6.7}$$

where W_t referred to the vector $\{x_i, y_{jk}\}$ evaluated at all times from 0 to t, and $A_i, B_{ki}, D_{ji}, E_{ji}$ were functionals with the constraint that all B_{ki} and all E_{ji} were nonnegative. He showed that if $C_i(t) = \theta_i C(t)$ was a spatial pattern input, his earlier limit theorems generalized to (6.8) under additional hypotheses on A_i, B_{ki}, C, D_{ji}, and E_{ji}. The heuristic meanings of these hypotheses, in the context of classical conditioning, were that each CS was presented sufficiently often for large times (that is, that the E_{ji} were "large" relative to the association decay rates D_{ji}), that the spatial pattern US was presented sufficiently often for large times (that is, that C was "large"

relative to $\Sigma_i X_i$), and that CS and US were practiced together sufficiently often for large times.

Since a spatial pattern could be either sensory or motor (for example), these equations described a network that could learn either to make associations with a complex sensory stimulus or to perform a complex motor act. The former occurs in classical conditioning and the latter in operant conditioning. Networks that obeyed (6.4) or (6.8) were used as subnetworks of larger networks that modeled classical and operant conditioning in a more realistic environment. Some of these larger networks will be discussed in Section 8.1.

6.3. *SYNAPSES WITH HABITUATION RULES*

The opposite of an associative rule for synaptic modification is a rule in which presynaptic activity followed by postsynaptic activity creates a tendency to *decrease* synaptic efficacy. Such a rule will be called *habituative*, even though "habituation" has a slightly different meaning for psychologists.

Grossberg [88, 89] developed a class of habituative synapses that he called *accumulation-depletion synapses*. He noted that while sufficiently repeated pairing between pre- and postsynaptic activity increased the *production* of chemical transmitter at the synapse, it also increased the *release* of transmitter. Thus, the net available transmitter (production minus release) could either increase or decrease with correlated activities. Such synapses have been used in models of a variety of psychological phenomena involving transient reactions to stimulus changes.

Grossberg fitted associative synapses and accumulation-depletion synapses into a unified theory in which the nature of the synaptic interactions depended on the sizes of specific parameters in one set of equations. To include both transmitter production and release in (6.4), it was necessary to replace the single variable $z_{jk}(t)$ by two variables: $z_{jk}(t)$ described the rate of transmitter production; $Z_{jk}(t)$ the amount of available transmitter. Thus, the function $z_{ki}(t)$ in (6.4a) was replaced by $Z_{ki}(t)$, and the equation for $Z_{jk}(t)$ was

$$\dot{Z}_{jk}(t) = \zeta_{jk}\{\eta_{jk} z_{jk}(t) - Z_{jk}(t)\} - \kappa_{jk}\left[x_j(t-\tau_{jk}) - \Gamma_{jk}\right]^+ Z_{jk}(t). \quad (6.8)$$

The three terms in (6.8) describe, respectively: transmitter production rate proportional to $z_{jk}(t)$, feedback inhibition proportional to amount of transmitter, and transmitter release proportional both to amount of transmitter and to presynaptic spiking frequency. If no transmitter release occurred, $Z_{jk}(t)$ was asymptotically proportional to $z_{jk}(t)$. If no learning occurred, the term $\eta_{jk} z_{jk}(t)$ in (6.8) was constant.

In [88] and [89], (6.8) with accumulation-depletion and no learning was used to construct a network that could learn a motor response that terminated

a painful stimulus such as electric shock. *Offset* of pain-related neural activity at one locus was transformed to transient *onset* of activity at another locus. The network is sketched in Figure 24 and will be described more fully in Section 8.1.

A different set of model synapses with habituation dynamics was developed by Uttley [205–209]. These synapses were described on the basis of classical information theory [191, 192].

Synapses with information-theoretic descriptions appeared earlier in Uttley [203], where the dynamics of plasticity were associative rather than habituative. This article proposed a linear threshold network with binary inputs x_i, $i = 1, \ldots, m$, and an output $(\sum_{i=1}^{m} \gamma_i x_i) + \gamma_0$, where the γ_i were synaptic conductivities and γ_0 was the negative of a response threshold (see Figure 14). The output cell responded if and only if

$$\left(\sum_{i=1}^{m} \gamma_i x_i \right) + \gamma_0 > 0, \tag{6.9}$$

and so became a discriminator of the class of patterns for which (6.9) holds. If $p(x)$ was defined as the average frequency of occurrence of the binary event x in some given time interval, the *mutual information* $I(x, y)$ between the two events x and y was defined (following Shannon) to be

$$I(x, y) = \log_2 \frac{p(x \cap y)}{p(x)p(y)}. \tag{6.10}$$

The conductivities γ_i were then set equal to $I(x_i, y)$ if $x_i = 1$ and $I(\bar{x}_i, y)$ if $x_i = 0$ ($\bar{x}_i$ being the complement of x_i). Uttley showed that a network of such elements could adapt to discriminate the most frequent pattern occurring in its input set.

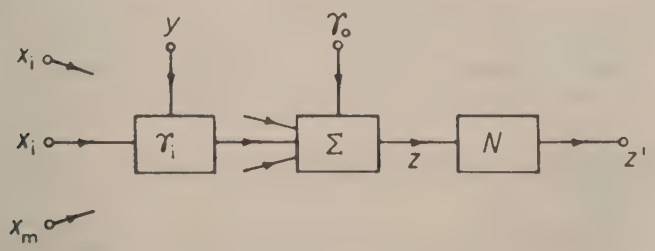

FIG. 14. Schematic linear pattern discrimination network. N denotes a threshold element, Z the total excitation it received. (From Uttley [203]; reprinted with permission from Elsevier Biomedical Press B.V.)

In [203] and [204], it was shown that a network whose synapses obeyed (6.9) and (6.10) was unstable because of the positive feedback loop between the synapse and the co-occurring patterns. The network thus had to be stabilized while preserving its pattern-discrimination properties. To do this, Uttley made conductivities be proportional to the *negative* of the information function. Thus, $\gamma_i = -k\log_2[p(x_i \cap y)/\{p(x_i)p(y)\}]$, and there was also a conductivity between "not x_i" and y, namely $\bar{\gamma}_i = -k\log_2[p(\bar{x} \cap y)/\{p(\bar{x}_i)p(y)\}]$. He gave the name *informon* to a network built on such conductivity laws.

In [205], the above equations were modified to include exponential smoothing over time of the frequencies. Defining $G(y)(t) = (1/T)\Sigma_{\tau=\infty}^{0} y(t)\exp(\tau/T)$ for any function $y(t)$ of integer time values and for a prespecified time interval T, the laws arrived at were

$$\gamma_i = -k\log_2 \frac{G(x \cap y)}{G(x)G(y)} \tag{6.11}$$

and

$$\bar{\gamma}_i = -k\log \frac{G(\bar{x} \cap y)}{G(\bar{x})G(y)}. \tag{6.12}$$

Changing the sign of the effect of cross-correlation on synaptic efficacy is not the only possible method of stabilizing a network with modifiable synapses. Other methods for bounding total activity include exponential decay with time of individual element activities and of synaptic efficacies (see Section 6.2), exponential decays combined with normalization of pattern weights (see Section 8.2), and conservation of total synaptic efficacy (also see Section 8.2). Uttley's method of negative-information synapses, however, has some interesting psychological consequences.

Uttley [206–209] used informon networks such as that of Figure 15 in models of classical conditioning. In [206] he studied, in particular, the case of a single CS-US connection. A CS, labeled A, excited a pathway of variable conductivity, and a US, labeled U, excited a pathway of fixed conductivity, both pathways in the same informon. The output of the informon was the CS. In this case, the steady-state network equations were shown to be

$$\gamma_A + \gamma_U \frac{G(U)}{1-G(U)} \left\{ \frac{G(A \cap U)}{G(A)G(U)} - 1 \right\} \approx 0. \tag{6.13}$$

If A was always reinforced, $G(A \cap U) = G(A)$, so (6.13) reduced to $\gamma_A + \gamma_U = 0$. Since $\gamma_A > 0$, because A developed a positive associational strength, this meant that $\gamma_U < 0$, that is, the pathway from U to the conditioned response (CR) was inhibitory.

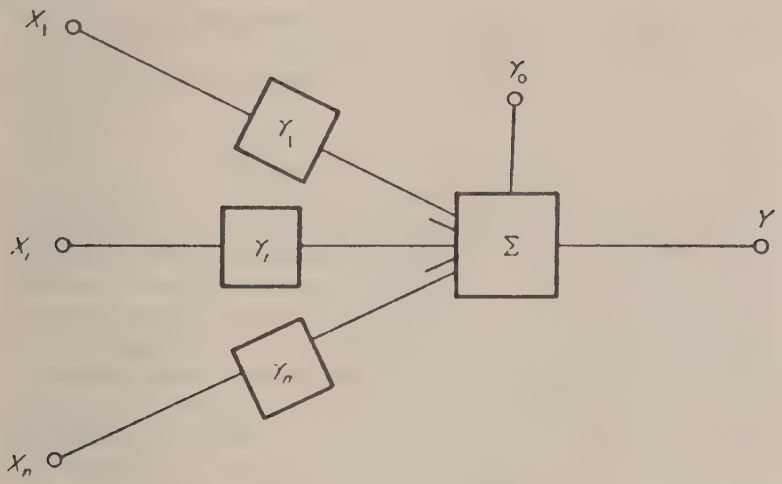

FIG. 15. Schematic linear informon network. (From Uttley [206]; reprinted with permission from Academic Press Inc. (London) Ltd.)

Uttley's theory could account for such classical conditioning data as the *blocking* and *overshadowing* paradigms [117]. In the blocking paradigm, the animal is first given many presentations of one CS (A), each followed by an electric shock as negatively reinforcing US. The CS A is then presented many times along with another stimulus X, each pair also followed by a shock. On recall trials, the animal reacts fearfully to A alone or to the AX combination, but not to X alone. In the overshadowing paradigm, shock is associated with the AX combination, but cue A is more *salient* than cue X—that is, either more intense or more important to the organism's survival. Again, no fearful reaction is learned to X alone.

The inhibitory US → CR connection, however, seems paradoxical because the US is a highly effective stimulus for evoking the unconditioned response (UR). Uttley justified the inhibitory connection by treating the conditioned and unconditioned responses as occurring in separate loci. Most contemporary psychologists, however, treat the CR as equivalent to a component of the UR.

The US → CR inhibition was further illuminated in [207, 208, 209]. These articles introduced an additional input called Z to the informon unit with a binary signal $F(Z)$ of fixed conductivity γ_z. The equation for the output signal became

$$F(Y) = \left(\sum_i F(x_i)\gamma_i\right) + F(Z)\gamma_z,$$

and the equation for change in conductivity with stimulus pairing became

$$\Delta\gamma_i = -kF(x_i)\Big(\sum_i F(x_j)\gamma_j + F(Z)\gamma_z\Big). \tag{6.14}$$

If there was to be a positive term in (6.14), γ_z had to be constant and *negative*.

Uttley called the inhibitory Z pathway a *classifying* pathway. He explained: "The function of $F(Z)$ must be to signal whether the total stimulus to all the variable pathways is a member or non-member of some class" [207, p. 28]. He discussed a computer simulation of a hierarchical network consisting of 210 informons in three layers, with a total of 8400 simultaneously variable pathways. This network discriminated handprinted numerals by looking for the occurrence or nonoccurrence of 50 primitive features. This discrete model of pattern discrimination, where a classification is equivalent to a subset of the set of features, is similar to those of Marr [140–142] discussed in Section 5. Continuous models of pattern discrimination will be discussed in Section 7.

Uttley [206] noted some formal analogies between his own model and that of Rescorla and Wagner [182], whose psychological inspiration was quite different. The Rescorla-Wagner model and related models will now be discussed.

6.4. *SYNAPSES WITH OTHER MODIFICATION RULES*

Rescorla and Wagner [182] developed a theory of the growth of CS-US associative strengths. Their theory was motivated by several psychological experiments which indicated that associative learning of a given CS can be greatly influenced by the background stimuli present during both training and recall trials. The cognitive basis of their theory was that "organisms only learn when events violate expectations. Certain expectations are built up about the events following a stimulus complex; expectations initiated by the complex and its component stimuli are then only modified when consequent events disagree with the composite expectation" [182, p. 75].

To provide a formal framework for this cognitive principle, Rescorla and Wagner proposed difference equations with variables defined in psychological rather than neural terms. Their variables were CS-US associative strengths as described in [115]. If V was the associative strength between a CS labeled A and a given US, then

$$\Delta V_A = \alpha_A \beta(\lambda - V_{AX}), \tag{6.15}$$

where α_A was the intensity of the CS, β is a learning rate associated with the given US, and λ was the asymptote of learning, which was a function of the

reinforcement strength of the US on the given trial. ($\lambda = 0$ if US reinforcement did not occur.) X referred to one or more stimuli present along with A, and it was assumed that the compound associative strength V_{AX} equaled $V_A + V_X$.

Rescorla and Wagner explained blocking (see Section 6.3) by noting that if A is strongly associated with shock, (6.16) predicts that $V_{AX} = V_A$ will become very close to λ. The analog of (6.16) for X is $\Delta V_X = \alpha_X(\lambda - V_{AX})$, and since at the time of recall trials V_{AX} is nearly equal to the value of λ for the given US, ΔV_X will be nearly 0.

Toward the end of their article, Rescorla and Wagner noted that blocking and overshadowing could be explained alternatively by a model based on selective attention. A more salient cue, or one that has already built up an association to the US, could receive more attention than a less salient cue and thereby inhibit learning of associations to the less salient cue. In Section 8, the two types of model will be compared in detail on the basis of further conditioning experiments.

The Rescorla-Wagner model of learning did not include specific neural interpretation for the variables used. Formal analogies exist, however, between (6.15) and other equations which have neural interpretations. For example, Uttley [206] noticed analogies between (6.15) and a version of his own informon equation. For a pair of conditioned stimuli A and B and an unconditioned stimulus U, Uttley derived from (6.11), (6.12), and a variant of (6.13) the difference equation

$$n_A \Delta\gamma_A = \frac{kn_A}{T\gamma_0}\left\{\frac{X_A}{G(X_A)} - 1\right\}\{X_U n_U \gamma_U + X_A n_A \gamma_A + X_B n_B \gamma_B\}, \quad (6.16)$$

with n_U, n_A, and n_B denoting numbers of pathways in the informon network. Uttley's overall conductivity $n_A\gamma_A$ corresponds to Rescorla and Wagner's associative strength V_A. The Rescorla-Wagner salience measure α_A corresponds to n_A, the number of pathways from CS A to the informon, while the learning rate B corresponds to Uttley's decay rate $k/T\gamma_0$. Finally, the asymptotic strength λ in (6.15) corresponds to the (negative) overall conductivity $n_U\lambda_U$ of pathways from the US to the informon.

The Rescorla-Wagner framework also shows analogies with a qualitative notion of synaptic change due to Klopf [125–127]. Klopf proposed that a given synapse was increased in efficacy if its activity was followed by a net *increase* in the depolarization received by the postsynaptic cell, that depolarization acting as a positive reinforcement signal. (Section 8.1 will discuss how this notion is embedded in Klopf's idea that both neurons and whole brains are goal-seeking devices, or "heterostats.")

Sutton and Barto [199] synthesized ideas from the Rescorla-Wagner and Klopf models. They based their classical conditioning model on n stimulus

traces $x_i(t)$, an output signal $y(t)$, and n CS-US synaptic weights $w_i(t)$. They also posited stimulus traces $\bar{x}_i(t)$ that were separate from the signals and longer lasting. Finally, they denoted by $\bar{y}(t)$ the weighted average of the values of y over some time interval preceding t. (The traces $\bar{x}_i$ were referred to as "eligibilities." Possible cellular mechanisms for these traces were suggested, involving calcium ions and cyclic nucleotides.) The values of these variables changed from time t to $t+1$ according to the equations

$$\begin{aligned} \bar{x}_i(t+1) &= \alpha \bar{x}_i(t) + x_i(t), \\ \bar{y}(t+1) &= \beta \bar{y}(t) + (1-\beta) y(t), \\ w_i(t+1) &= w_i(t) + c\{y(t) - \bar{y}(t)\}\bar{x}_i(t), \\ y(t) &= f\left(\sum_{j=1}^{n} w_j(t) x_j(t) \right), \end{aligned} \tag{6.17}$$

where α, β were constants ($0 \leqslant \alpha, \beta < 1$), f was a sigmoid function, and $c > 0$ was a constant determining the rate of learning.

The differential-equation approximation of (6.17c) would be an equation for an associative synapse except that $y - \bar{y}$ replaced y alone. Thus, Sutton and Barto were consistent with Klopf's idea that efficacy was influenced by an increase in the postsynaptic depolarization rather than by the total amount of depolarization. They further noted that the learning rate constant c in (6.17) was analogous to the product of α_A and β in (6.15). Also, the term $y(t)$ was analogous to λ in (6.15), both denoting received US values. Sutton and Barto state, in reference to Rescorla and Wagner: "Whereas in that model the associative strengths are changed based on the difference between received UCS levels, in our model weights are changed based on the difference between actual activity level y and expected activity level $\bar{y}$" [199, p. 149].

The utility of different types of modifiable synapses as elements in large-scale brain models will be considered in Section 8.1. In particular, models of conditioning involving synapses with rules like those of this section will be compared with models involving associative synapses.

6.5. *GENERAL REMARKS ON MODIFIABLE SYNAPSES*

The biochemistry of synapses and postsynaptic receptors is an active area of current experimental work. Although such substances as acetylcholine, norepinephrine, and gamma-amino butyric acid have long been known to be neural transmitters, little is known conclusively about how those substances affect ionic conductances or how the transmitters, in turn, are affected by pre- and postsynaptic electrical activities. A popular, though not yet established, idea is that some transmitters act through a "second messenger" such

as a cyclic nucleotide (cf. [71, 217]). Further experimental results in this area will indicate which synaptic models are simultaneously plausible at the cellular and organismic levels.

While the types of plastic synapses that exist are not established, *some* kind of plasticity is a parsimonious hypothesis for long-term memory. Since Hebb [109], it has been believed that short-term memory involves different mechanisms, perhaps including hysteresis in a feedback network or loop (cf. Section 4). (It is worth noting, however, that some evidence exists suggesting synaptic plasticity can also mediate short-term habituation in invertebrates [118, pp. 23–37].) The next section will describe short-term memory and related effects in lateral inhibitory networks with nonmodifiable synapses. Section 8 will discuss models which synthesize the mechanisms of short-term and long-term memory.

7. PATTERN PROCESSING BY NETWORKS WITH LATERAL INHIBITION

The idea of lateral inhibition, or on-center off-surround, in sensory pattern processing (see Figure 16) dates back at least to the last century. In their theories of visual perception, both Helmholtz and Mach explained the selective enhancement of edges or contours by means of networks of retinal cells each excited by light within a central area and inhibited by light within a surrounding area. There is by now much experimental evidence that subsystems exist in many areas of the central and peripheral nervous system where inhibition operates over a greater distance than excitation. The evidence began with the classical studies of the compound eye of the horseshoe crab (*Limulus*) by Hartline and Ratliff [108] and of the vertebrate retina by

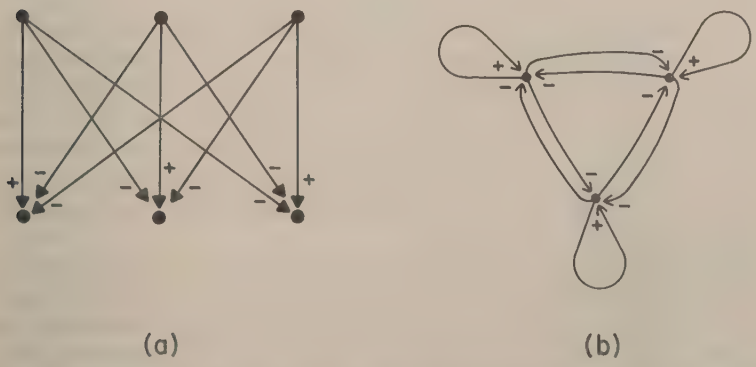

FIG. 16. Examples of two kinds of lateral inhibitory networks: (a) nonrecurrent (feed-forward); (b) recurrent (feedback).

Kuffler [132]. The organization of some areas of the sensory cortex into columns of cells with lateral interactions between them [113, 114, 154] has led many investigators to believe that the on-center off-surround principle also operates at the neocortical level. Evidence for lateral inhibition in other areas such as the cerebellum and the hippocampus is summarized in [90].

In recent years, lateral inhibition has often been an organizing principle in model neural networks, particularly networks involved in pattern discrimination. Figure 16 schematizes the difference between two general types of lateral inhibition: *nonrecurrent* (feed-forward) and *recurrent* (feedback) inhibition. Other classifications of lateral inhibitory networks include multiplicative (or shunting) versus additive interactions, and lumped interactions (excitatory and inhibitory cells treated as part of the same population) versus unlumped interactions (excitatories and inhibitories treated separately).

7.1. *EARLY STUDIES AND GENERAL THEMES (CONTRAST, ENHANCEMENT, NORMALIZATION, COMPETITION)*

In 1957, Hartline and Ratliff [108] modeled inhibition in the *Limulus* eye via a pair of simultaneous algebraic equations for two mutually inhibiting receptors, as follows:

$$\begin{aligned} r_1 &= \left(e_1 - k_{1,2}\left(r_2 - r_{1,2}^0\right)^+\right)^+, \\ r_2 &= \left(e_2 - k_{2,1}\left(r_1 - r_{2,1}^0\right)^+\right)^+. \end{aligned} \tag{7.1}$$

Here r_i denotes the impulse frequency in the axon of cell i, and e_i denotes the excitation supplied by an external stimulus. $r_{1,2}^0$ and $r_{2,1}^0$ were threshold frequencies that each cell had to exceed before it could exert inhibition, and $k_{1,2}$ and $k_{2,1}$ were coefficients of inhibitory action.

The Hartline-Ratliff equations proved effective in the modeling of a variety of experimental data, and other early lateral inhibitory models used and elaborated them. For example, Reichardt and MacGinitie [181] discussed a nonlinear system of equations for recurrent lateral interactions in the *Limulus* eye. Logarithmic transduction of light inputs was followed by a linear threshold network with inhibitory coefficients. The stability of network voltages was considered for various parameter choices, and applications to pattern recognition were discussed.

Sperling and Sondhi [195] developed a lateral inhibitory model of some of the more complex circuitry of the mammalian retina. They used three stages of an electrical network—low-pass filtering, feedback, and feed-forward—to explain certain data on luminance and flicker detection. The feedback stage, corresponding to retinal feedback from horizontal cells to photoreceptors,

was described by the equations

$$\frac{d}{dj}v_j(\lambda)+v_j(\lambda)\left[1+v_m(\lambda)\right]=v_{j-1}(\lambda), \qquad j=1,\ldots,n. \tag{7.2}$$

The nonlinear term $v_j(\lambda)v_n(\lambda)$ in (7.2) corresponded to multiplicative (shunting) inhibition exerted by the nth stage of cells on the jth stage. Such shunting inhibition (and shunting excitation) were described by Hodgkin [111] in connection with the effects of a presynaptic neuron on the various postsynaptic ionic conductances. The major effect of this shunting inhibition was described in [195] as "reducing dynamic range." Similar effects have often been described in lateral inhibitory networks under various names: "contour enhancement" [90, 180], "input-pattern sharpening" [151], "crispening of pattern weights that can fire a given cell" [87], and "contrast enhancement" [56, 103].

Contrast enhancement is a form of decision or competition between inputs, and similar decision mechanisms can exist at many levels of the brain. The work of Kilmer et al. [122] on decisions between competing modes in the reticular formation, for example, was discussed in Section 5; Grossberg [84] noted that this property could be formally achieved in a nonrecurrent on-center off-surround network. Competition between inputs, particularly at the level of the sensory cortex, will be a major theme of later sections of this chapter.

On-center off-surround networks were also used in [84] to model the vertebrate retina. A mechanism was developed for making a cell fire in response to one and only one space-time pattern—that is, to a particular distribution of inputs as it varies over time. This mechanism is shown in Figure 17. The first stage involved a nonrecurrent lateral inhibitory network for low-band filtering. The second stage involved a different mechanism for high-band filtering (switch from net excitation to net inhibition with increase in spiking frequency). The nonrecurrent lateral inhibitory stage also accomplished *pattern normalization*, which is schematized in Figure 18. In other words, population activities $x_1,\ldots,x_n$ at one ("input") level are replaced by activities $y_1,\ldots,y_n$ at the next ("output") level which were proportional to the x_i's but independent of the total input energy $\Sigma_i x_i$. Normalization has often been used in lateral inhibitory networks to keep the total network activity bounded.

Grossberg [87] compared the above two stages of inhibition to the horizontal and amacrine cell layers in the retina. These retinal mechanisms were integrated with a (cerebellar) expectation mechanism (see Section 5.2) to build a network that discriminates a space-time pattern changeable by conditioning.

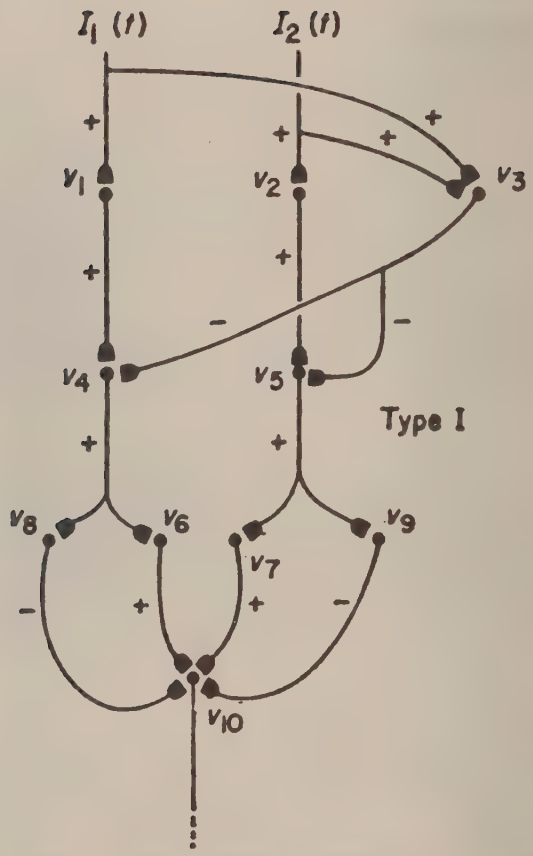

FIG. 17. Network for discriminating a single space-time pattern. The subnetwork v_1 through v_5 accomplished low-band filtering and pattern normalization (see next figure). The subnetwork v_4 through v_{10} accomplished high-band filtering. (From Grossberg [84]; reprinted with permission from Academic Press Inc. (London) Ltd.)

In [84] and [87], nonrecurrent lateral inhibition was chosen rather than recurrent lateral inhibition because a brief representation of the sensory pattern was advantageous. For the opposite reason, keeping patterns alive in short-term memory, *recurrent* lateral inhibition tends to be preferred for models of pattern processing at the cortical level [90, 214].

Another article by Grossberg [78] added some lateral inhibition terms to equations similar to those discussed in Section 6.2. This was done to facilitate the learning of an association such as AB if a competing association such as AC was being presented in the same time period but less often than AB. He

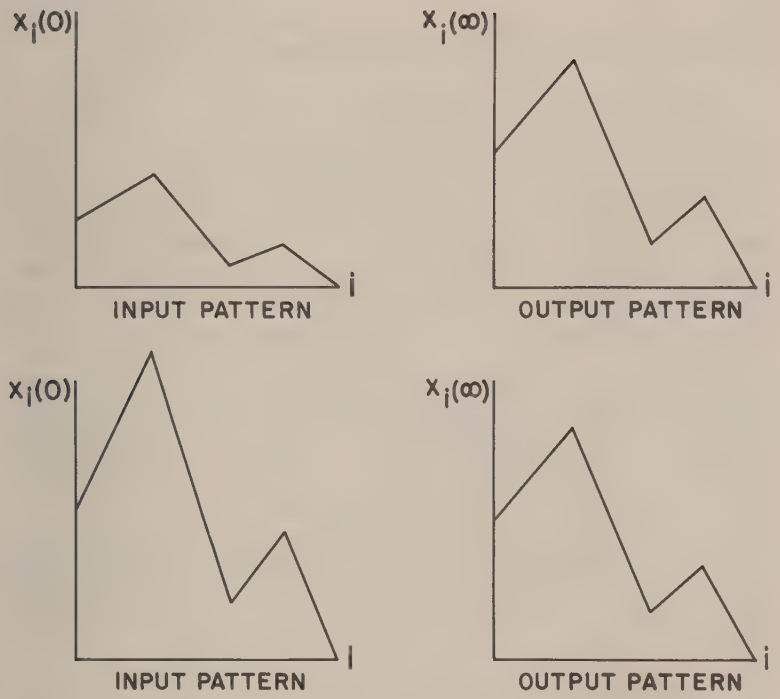

FIG. 18. Example of pattern normalization. Output pattern has same *relative* activities as input pattern, but is independent of *absolute* input activity.

noted that (7.1) could be derived from the steady-state solution of his own equations.

Morishita and Yajima [151] modeled n mutually inhibiting neurons with nonnegative outside inputs. The equations they derived were markedly similar to a nonlinear variant of (8.1), and were also a special case of Equation (1) of [84]. In [151], either single or multiple equilibria were possible, and hysteresis could occur between multiple equilibria if any existed. The single-equilibrium case corresponded to contrast enhancement. In one subcase of the multiple-equilibrium case, where all connection weights were equal, there was a one-to-one correspondence between the input and the equilibrium reached at time ∞, corresponding to faithful information storage. In another subcase, where each neuron was inhibited only by its two nearest neighbors, the system could become contrast-enhanced, or else enter periodic orbits.

Concurrently with these models, recurrent lateral inhibitory networks were being developed for interactions in the sensory cortex. Such networks included

both inhibition and excitation, and had stochastic effects at the neuronal level that were averaged out at the population level.

7.2. *LATERAL INHIBITION AND EXCITATION BETWEEN POPULATIONS*

One series of articles on models of lateral inhibition in the sensory (particularly visual) cortex includes Wilson and Cowan [214, 215] and Ermentrout and Cowan [57]. In [214], the elements of the network were not single neurons, but neuron populations in which intrapopulation connections were random and interpopulation connections were deterministic. Step functions with a distribution of different cell thresholds were averaged out into sigmoid functions (see Figure 2) describing input-output relations at the population level.

The model of [214] included excitation as well as inhibition between populations. The excitation tended to fall off more sharply with distance than the inhibition. Refractory periods were included, leading to time-integral terms that were then removed by averaging the population activities over a suitable time interval. Thus, the network could be described in terms of the two functions $E(t)$ and $I(t)$ (proportions of excitatory and inhibitory cells, respectively, firing per unit time at time t). If $P(t)$ and $Q(t)$ were the inputs to excitatory and inhibitory populations, respectively, the resulting equations were of the form

$$\begin{aligned} \tau_e \frac{dE}{dt} &= -E + (k_e - r_e E)\mathcal{S}_e(c_1 E - c_2 I + P), \\ \tau_i \frac{dI}{dt} &= -I + (k_i - r_i I)\mathcal{S}_i(c_3 E - c_4 I + Q), \end{aligned} \tag{7.3}$$

where $\mathcal{S}_e$ and $\mathcal{S}_i$ were sigmoid functions. It was noted in [90] that the coefficients $k_e - r_e E$ and $k_i - r_i I$ were equivalent to the terms for shunting (multiplicative) excitation and inhibition in the single-neuron passive membrane equations.

Wilson and Cowan found that (7.3) had either one, three, or five steady states, and could exhibit either limit-cycle behavior or simple or multiple hysteresis (see Figure 19). In their Theorem 4, they showed by local eigenvalue analysis that "any neural population which exhibits limit cycle activity for some class of stimulus configurations will also display simple hysteresis phenomena for some other class of stimulus configurations" [214, p. 20]. Their physiological interpretation was: "The theorem shows that nonspecific biasing inputs to a neural population from other parts of the central nervous system may completely change the character of the response of that population to specific sensory (or experimental) stimulation" [214, p. 20].

A second article by Wilson and Cowan [215] described a two-dimensional network for representing an area of cortex or thalamus. This network had

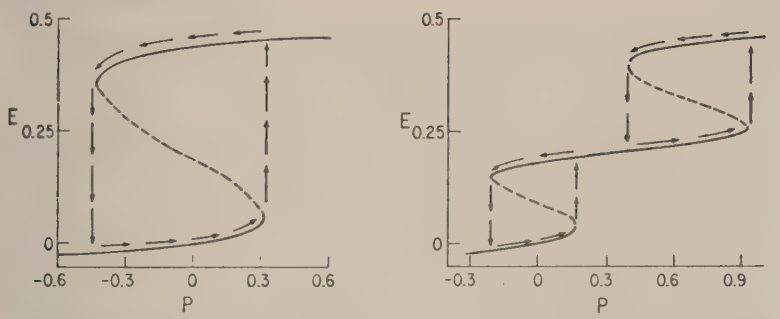

FIG. 19. Steady-state value of E as function of P ($Q = 0$) for (7.4), for two different sets of parameters. Hysteresis loop indicated by arrows occurred if P was increased and then decreased. Dotted lines were unstable states. Sigmoids used were of form $S(x) = 1/(1 + \exp[-\alpha(x - \theta)])$. (From Wilson and Cowan [214]; reprinted with permission from The Biophysical Society.)

properties similar to that of [214] with the addition of distance-dependent interactions. The resulting equations were integrodifferential equations that had the same range of limiting behavior as before, including multiple hysteresis and limit cycles. The limit cycles were seen as possible analogs of the reverberatory thalamocortical loops often postulated for short-term memory (cf. [13]). This network modeled various phenomena of visual psychophysics, "including spatial modulation transfer function determinations, certain metacontrast effects, and the spatial hysteresis phenomenon found in stereopsis" [215, p. 55].

Wilson has recently developed a variety of more detailed models of visual phenomena (for example, [212, 213]). These models have, in general, used sigmoid functions, but not an on-center off-surround anatomy. The model of [212] for spatial frequency adaptation will be discussed in Section 8.2.

Ermentrout and Cowan [57] proved bifurcation results for a model network that was partly inspired by the Wilson-Cowan networks. Distance-dependent interactions were included, but, for the sake of mathematical tractability, shunting interactions were entirely removed. The system of ordinary differential equations arrived at was

$$\mu \frac{dV_i}{dt} = -V_i + \sum_k \alpha_{kj} S(V_k(t)) + P_i(t), \qquad i = 0,1,2,\ldots,$$

with the weight α_{kj} taken to be a function of $|j - k|$. If the tissue was assumed to be isotropic, the neurons could be taken in the limiting case to be

a continuum, leading to the vector integrodifferential system

$$\frac{\partial \mathbf{V}}{\partial t} = -\mathbf{V} + \mathbf{K} * \mathbf{S}(\mathbf{V}) + \mathbf{P},$$

with $\mathbf{K}$ a kernel and $*$ denoting spatial convolution. Finally, defining $\mathbf{Y}$ by $\mathbf{V} = \mathbf{K} * \mathbf{Y} + \mathbf{P}$ and considering the steady state led to the system

$$-\dot{\mathbf{Y}} + \mathbf{S}(\mathbf{K} * \mathbf{Y} + \mathbf{P}) = 0. \tag{7.4}$$

Bifurcating spatially periodic solutions were found for (7.4) and related equations under a variety of conditions. The authors state: "The biological significance of these spatially periodic solutions is not completely clear although they may be related to certain simple visual hallucinations" [57, p. 15].

Another series of articles on lateral inhibition between cortical populations, going from 1973 to the present, includes Grossberg [90, 97], Grossberg and Levine [103], Ellias and Grossberg [56], Levine and Grossberg [137], Levine [135, 136], and Cohen and Grossberg [48]. In many versions of the lumped case, where each population was both excitatory and inhibitory, the activity of the ith population was given by

$$\dot{x}_i = -Ax_i + (B_i - x_i) \sum_{k=1}^{n} f(x_k)\gamma_{ik} - (x_i - C_i) \sum_{k=1}^{n} f(x_k)\phi_{ik} + I_k(t), \tag{7.5}$$

where f was an increasing function with $f(0) = 0$, B_i and C_i were the maximum and minimum possible activities of the ith population, γ_{ik} and ϕ_{ik} were excitatory and inhibitory interaction coefficients, and $I_i(t)$ was the input to the ith population. The *signal function* f reflected neuronal input-output transformations.

Grossberg's equation (7.5) has features in common with the Wilson-Cowan equations (7.3). A major difference between the two sets of equations, however, was discussed in Section V of [90]. In (7.5), the sigmoid (or other monotone increasing) function of the inputs was computed separately for the excitatory and inhibitory inputs. Therefore, the excitatory and inhibitory inputs combined not additively as in (7.3), but via shunting terms.

Grossberg [90] argued that the separate inhibitory and excitatory "gain control" allowed for automatic retuning of network sensitivity in response to fluctuating inputs. He also argued that his own equations reflected an averaging on the population level of the passive membrane equations for a single neuron. B_i and C_i could be interpreted on the single-cell level as the

equilibrium potentials for sodium (Na^+) ions and potassium (K^+) ions respectively, and the factors $\Sigma_{k=1}^n f(x_k)\gamma_{ik}$ and $\Sigma_{k=1}^n f(x_k)\phi_{ik}$ as conductances of those ions. In his words, "a marriage between two experimentally verified phenomena—passive membrane equations and recurrent on-center off-surround anatomies—is hereby demonstrated" [90, p. 244].

Global theorems were proved on the limiting behavior as time approached infinity of various subcases of (7.5), a generalization of (7.5), and its unlumped modification. Computer simulations showed that the network quickly reached activities close to its limiting values (in a few minutes if decay rates were chosen to reflect neuronal time constants). The limiting behavior, therefore, approximated finite-time sensory pattern transformations such as those occurring in short-term memory.

In [90], the case where the maximum activities B_i were all equal was studied. The values $x_i(0)$ reflected the input pattern, and pattern transformations after time 0 reflected the recurrent interactions. The minimum activities C_i were set to 0. The connectivity of the network was pure on-center off-surround [see Figure 16(b)]; that is, the excitatory coefficients γ_{ik} were set equal to δ_{ik} (that is, 1 if $i = k$, 0 if $i \neq k$), and the inhibitory coefficients ϕ_{ik} to $1 - \delta_{ik}$.

Grossberg showed that for broad classes of monotone increasing functions f, the variables x_i in (7.5) always converged to limits. The proofs used global analyses via monotonicity methods. These limits were interpreted as meaning that the network arrived at a choice as to what parts, if any, of a pattern were to be stored in short-term memory. The choice made depended on the function f. If f was linear in x, the output values $x_i(\infty)$ were proportional to the input values $x_i(0)$; in the case of a visual pattern, for example, relative reflectances were conserved. Although proportionality would seem to be a good property for a short-term memory system, in reality it was not, because insignificant network noise could be stored as well as significant signal traces. A better outcome, Grossberg stated, was for $x_i(\infty)$ to be proportional to $x_i(0)$ if $x_i(0)$ was above a threshold value (*quenching threshold*) and equal to 0 if $x_i(0)$ was below that threshold (contrast enhancement with noise suppression). He found that the desirable limiting behavior could be achieved if a sigmoid function f was used (cf. Figure 2).

In [103], many of these results were generalized to (7.5) with unequal B_i. This meant that the competition between populations, for storage of their preferred sensory features, was biased in favor of some populations. Possible sources of such a bias were: (1) that some stimuli occurred more often during development than others, causing the unequal growth of relevant feature detectors (cf. [40, 110]; (2) that some stimuli were attended to more than others during adult life; or (3) that the network contained statistical errors. Monotonicity and Liapunov arguments were used to show that global limits existed for broad classes of functions f. The result of [90] on contrast

enhancement with noise suppression still held within a *subfield* (defined as the subnetwork of all populations with a given value of B). A choice was made *between* subfields, and the asymptotic activities of all but a small number of subfields were 0 as t went to infinity. The subfields chosen were the result of a "tug-of-war" between the subfields with largest inputs and the subfields with largest maximum activities. Figure 20 shows pattern transformations over time in such a network.

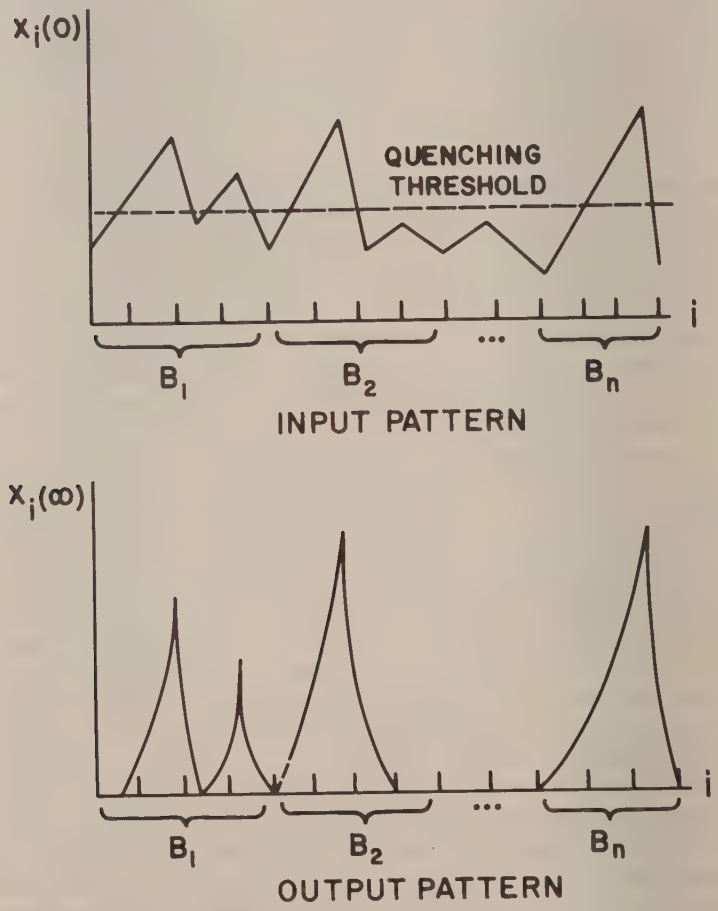

FIG. 20. Examples of pattern transformation in the network of Grossberg and Levine [103], where $B_n > B_{n-1} > \cdots > B_1$, with $f(x)$ a properly chosen sigmoid. Contrast enhancement with noise suppression held in subfields with nonzero asymptotic activities. (Reprinted with permission from Academic Press Inc. (London) Ltd.)

These limit theorems were further generalized in [48], [95], [97], and [136]. In particular, it was shown in [97] that global limits existed for the competitive system

$$\dot{x}_i = A_i(\mathbf{x})\big[B_i(x_i) - C(\mathbf{x})\big], \qquad i = 1,\ldots,n, \qquad \mathbf{x} = (x_1, x_2, \ldots x_n), \tag{7.6}$$

with the A_i nonnegative, the B_i bounded, and C nondecreasing with respect to each x_i. Equation (7.6) generalized the subcase of (7.5) with $\gamma_{ik} = \delta_{ik}$, $\phi_{ik} = 1$. It also included cases where different populations had different signal functions. In [48], the existence of limits was generalized to some systems with distance-dependent inhibition ($\gamma_{ik} = \delta_{ik}$ but arbitrary ϕ_{ik} as long as $\phi_{ik} = \phi_{ki}$). That proof was based on a Liapunov-function method combined with Sard's theorem for manifolds.

In [137], some human visual illusions were simulated numerically using (7.5). In these simulations, each population was interpreted as a column of visual cortical neurons sensitive to a given line orientation (cf. [113, 114]). The interaction coefficients decreased with increasing distance between populations, but inhibition was broader than excitation; that is, γ_{ik} decreased faster than ϕ_{ik}. Fixed stimuli were given, denoted by nonzero inputs to a few populations. If all B_i were equal, the network simulated the result [39] that an acute angle is seen as slightly larger than it really is. This occurred because lateral inhibition by one population shifted the local maximum of activity from another, simultaneously active population to a nearby one. If the populations responsive to vertical or horizontal orientations had larger B_i than others, the network simulated the result [69] that a fixed near-vertical tilted line appears closer to vertical after prolonged viewing. In [135], a numerical study was done of (7.5) in which the populations coded visual-field position instead of orientation; this network simulated the binocular hysteresis date of [59] (cf. [215]).

In [56], various unlumped networks (see Figure 21) were studied in which no competitive biases existed. The equations for one example were

$$\left.\begin{aligned} \dot{x}_i &= -Ax_i + (B - X_i)f(x_i) - x_i \sum_{k \neq i} g(y_k) \\ \dot{y}_i &= -Ey_i + h(x_i) \end{aligned}\right\} \qquad i = 1,\ldots,n, \tag{7.7}$$

where x_i were excitatory activities, y_i were inhibitory activities, and f, g, and h were monotone increasing functions. As E became large, (7.7) became asymptotically equivalent to a subcase of the lumped system

$$\dot{x}_i = -Ax_i + (B - x_i)f(x_i) - x_i \sum_{k \neq i} g(x_k) + I_i, \qquad i = 1,\ldots,n \tag{7.8}$$

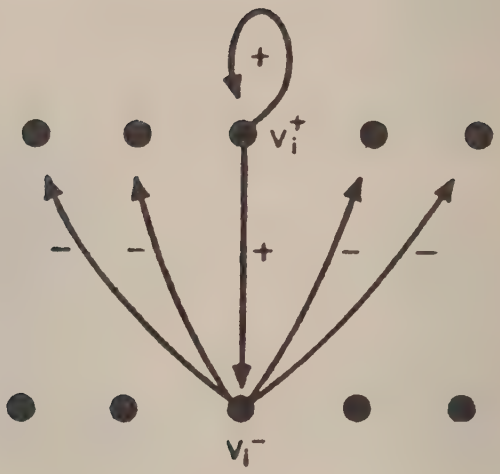

FIG. 21. Example of unlumped recurrent on-center off-surround anatomy. (From Ellias and Grossberg [56]; reprinted with permission from Springer–Verlag.)

which was like (7.5) except that excitatory and inhibitory signal functions were different. For both (7.7) and (7.8), global limits existed, but how much contrast enhancement occurred depended on the ratio of inhibitory to excitatory gain. In (7.7), if inhibition was slow relative to excitation (that is, if E was small), the limiting pattern could either have all weights equal to $1/n$, zero weights in all but the one population which had maximal activity at $t = 0$, or equal nonzero weights in a subfield of populations which had maximal activity at $t = 0$. The form of the limiting distribution depended on the signs of two quantities which included network-gain terms.

Slow inhibition in an unlumped network could also lead to pattern processing unlike that occurring in an asymptotically lumped network. An example was the two-dimensional system

$$\begin{aligned} \dot{x} &= -Ax + (B - x)[Cf(x) + I] - Dxf(y), \\ \dot{y} &= E(x - y) \end{aligned} \tag{7.9}$$

for excitatory and inhibitory activities in an unlumped network with uniform initial data and uniform inputs. Analytical and numerical studies of (7.9), including Hopf bifurcation, were done as the input I was varied. In one typical example, the asymptotic behavior moved from a stable equilibrium to a stable limit cycle and back as I increased (see Figure 22). When (7.9) was modified to allow for nonuniform initial data, stable limit cycles still oc-

curred, but did not preserve the initial order of pattern weights. It was concluded that slow, periodic waves such as are commonly seen in real brains [13] were unlikely to store the patterns in STM. Such periodic waves could, however, be "acting as shunts that continually retune the structures that are capable of STM storage" [56, p. 92].

The results of this section indicate that environmental changes can alter the processing carried out by an on-center off-surround network without altering the overall network structure. While the network is relatively insensitive to short-term input fluctuations, its parameters can be changed by

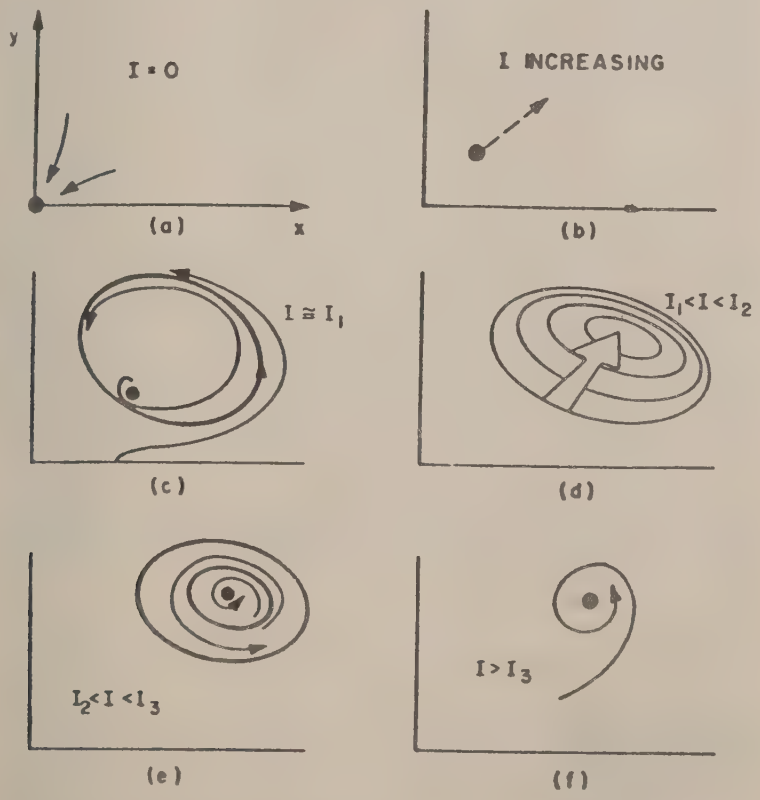

FIG. 22. Observed sequence of phase portraits when uniform input I is increased in the unlumped model of (7.10). (From Ellias and Grossberg [56]; reprinted with permission from Springer–Verlag.)

long-term environmental shifts such as those occurring with novel stimuli or with changing reinforcement criteria. Such changes in the equation parameters operate over a time scale that is long relative to the operation modeled by the equations, pointing to possible mechanisms for the interactions between short-term and long-term memory. Such STM-LTM interaction will be discussed in Section 8.2.

7.3. *SYSTEMS WITH COMPETITION AND COOPERATION*

Many researchers have noted that the competition between sensory features in lateral inhibitory neural networks has formal analogies with competition in other biological systems, such as animal populations in ecology or chemical species in embryology (see, for example, [95, 97]). Some recent studies have carried this analogy further by developing networks that combine competition in one dimension with cooperation in another dimension.

Competitive and cooperative elements were combined in the S-RETIC simulation of Kilmer et al. [122] discussed in Section 5. The competition was between different modes of behavior; the cooperation was between different modules that received different sensory information. Montalvo [150] noted the structural similarity (see Figure 20) between this mode-decision network and the network developed by Dev [51] to model stereoscopic vision. In the Dev model, the "competition dimension" was binocular disparity, and the "cooperation dimension" was retinal-field position. The network reached a consensus as to overall depth by correlating disparity information at different points in a complex binocular pattern. [Montalvo extended the analogy to the process (Didday [52]) by which the frog tectum determines a snapping position. However, the Didday model contained only competition and no cooperation.]

The Kilmer et al. and Dev models led to unique stable equilibria for a given input pattern. Amari and Arbib [10] developed competitive-cooperative models that included the possibility of multiple stable equilibria (and, therefore, hysteresis). Their equations were

$$\frac{\partial u(x,d,t)}{\partial t} = u(x,d,t) + \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} w_1(x-\xi, d-\eta) f[u(\xi,\eta,t)]\, d\xi\, d\eta - \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} u_2(x-\xi) g[v(\xi,t)]\, d\xi - h_1 + s(x,d,t)$$

$$\tau \frac{\partial v(x,t)}{\partial t} = -v(x,t) - h_2 \int_{-\infty}^{\infty} f[u(x,\eta,t)]\, d\eta,$$

where u and v were the excitatory and inhibitory activities, respectively; x and d were the cooperation and competition dimensions, respectively; w_1 and

w_2 were weighting functions; and f and g were monotone nondecreasing input-output functions. The spread w_1 was assumed to be broad in the x direction and narrow in the d direction; the spread of w_2 was assumed to be narrow. It was the spread of excitation that produced multiple stable steady states, even in the simplified model where no cooperation existed.

The theory of networks combining competition and cooperation is an area of much current research (see, for example, the 1982 conference proceedings in that area [11]). Cooperation has been particularly prominent in recent models of stereo vision (for example, Marr and Poggio [144, 145]). Grossberg [102] argued that a competitive network, such as those discussed in Section 7.2, can account for some of the major phenomena in stereo vision without the need for a separate cooperative dimension. However, many of the models of Grossberg and his coworkers (for example, [56] and [137]) involve "cooperation" within a competitive field, in the form of narrow-range recurrent excitation. The question of whether the extra dimension is needed remains an open one, as is the possibility of extending this type of theory to other neural processes.

7.4. *GENERAL REMARKS ON ON-CENTER OFF-SURROUND ANATOMIES*

A number of coherent results have emerged in the different models using on-center off-surround anatomies. For instance, there is widespread agreement among modelers that one function of lateral inhibition is to sharpen the sensory patterns that impinge on the network. There is also widespread agreement on the functions of nonrecurrent inhibition (low-level processing and simple decisionmaking) versus recurrent lateral inhibition (short-term memory storage). Distinctions between different on-center off-surround models are on the level of deeper organizing principles that influence *which* decisions are made or which patterns are stored. The organization of decision-making, in turn, influences how well such a network can be imbedded in a larger network involved in complex behavioral sequences. The next chapter will explore how such larger networks could be organized.

8. INTERACTION BETWEEN SUBNETWORKS AT DIFFERENT LEVELS: TOWARD GLOBAL BRAIN MODELS

Ever since perceptrons were created with sensory, association, and response units [184], all serious efforts at modeling global brain organization have treated the brain as made up of distinct functional subnetworks. The subassembly outlook is shared by both neural modelers and artificial-intelligence theorists (cf. [193, pp. 90–93]).

Most brain modelers find some merit in the slogan that "the brain is an action-oriented layered somatotopically organized computer" [24]. More needs to be said, however, on the details of how different levels of the brain

interact. Complex psychological processes like conditioning, reinforcement, and sensory coding involve the integration of pattern processing with learning and with motivation. Theories of these processes must therefore integrate models of the type described in Section 6 with other models of the type described in Section 7.

8.1. *CONDITIONING, REINFORCEMENT, AND SELECTIVE ATTENTION*

A series of articles from 1971 to 1975 by Grossberg [85, 88, 89, 91] discusses successive refinements of a network model of conditioning. These networks were built using the *method of minimal anatomies*:

> The theory tries to isolate postulates that act as guiding principles of neural design during individual development and the evolution of species. The networks that are hereby derived are capable of behavior that is far more complex and subtle than the postulates themselves, and also generate various new predictions. The theory is derived by a method of successive approximations; as more postulates are imposed, the networks become ever more sophisticated and realistic. At each stage of the derivation, basic mechanisms of network organization emerge, and are preserved as new postulates are imposed. [91, p. 264]

In [85], a unified mechanism for both classical and operant conditioning was built, using modifiable synapses that obeyed (6.7). That mechanism involved sensory loci (called $\mathcal{S}$) that were responsive to the CS, together with motor or autonomic loci (called $\mathcal{M}$) that were active after the US was presented. The fundamental postulate was: if a CS repeatedly precedes a US, the CS must learn to generate a motor response (UR) previously associated with the US. The problem was that the CS-US time lag could be variable. Thus, if the network was not designed carefully, the CS could become associated not with the US but with a noisy mixture of other patterns experienced in the same time period. The network that solved this problem is reproduced in Figure 23. The task was facilitated by including a third set of loci, called $\mathcal{A}$ for arousal, separate from $\mathcal{S}$ or $\mathcal{M}$. The network allowed repeated CS-US pairing to strengthen the ability of the CS to activate an arousal locus already strongly connected to the US.

Some of the $\mathcal{A}$ loci were representations of the various drives (hunger, sex, and so on). These loci received inputs from $\mathcal{S}$, via modifiable synapses, and from a nonspecific arousal source. Analogies with brain areas were discussed in Section 5.2. To keep the CS representation alive in short-term memory long enough for the conditioned association to develop, a recurrent on-center off-surround anatomy was used (see Section 7.3).

In [88] and [89], the new question was: How can an operant response that terminates a painful stimulus (for example, electric shock) become positively reinforced? For that to occur, the offset of pain-related neural activity had to be transformed into the transient onset of excitatory neural activity at

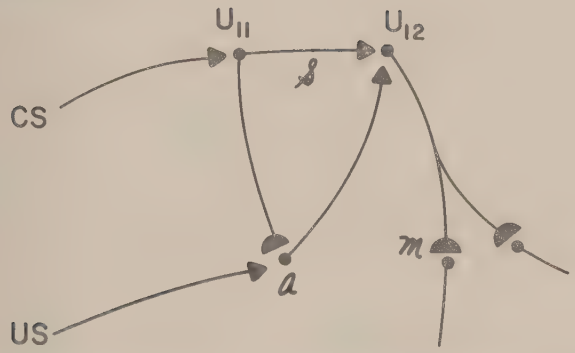

FIG. 23. Network for implementing classical conditioning; variant also implemented operant conditioning. Semicircles denote associative synapses; arrows denote unmodifiable synapses. (Modified from Grossberg [85]; printed with permission from Academic Press Inc. (London) Ltd.)

another locus. After that transformation occurred, motor responses that led to shock termination became reinforced in the same way that responses leading to hunger satisfaction were reinforced in [85].

This network, shown in Figure 24, involved modifiable synapses of the accumulation-depletion type (see Section 6.3). There were two competing channels, one for negative and one for positive reinforcement; both received a steady (tonic) arousal input. While shock was on, the "negative" channel received an additional input, making that channel have greater activity than the "positive" one. At the offset of shock, however, the "negative" channel was more depleted of transmitter than the "positive" channel. Thus, the combination of tonic arousal and lesser depletion gave the "positive" channel greater activity.

Grossberg [89] suggested a testable measure for the reinforcement value of a reduction from I to J units of shock for given values of I and J. Also, he generalized the rebound for negative to positive to a rebound from positive to negative, providing a mechanism for the frustration and counterconditioning that arises when expected reinforcement is not delivered. This notion of *gated dipole* (positive versus negative) became a major organizing principle of his more recent work, including models of such visual phenomena as aftereffects and spatial frequency adaptation [99].

Grossberg [91] refined the network further to model some data on attention. First, he noted that in the network of Figure 23, nonspecific arousal made it possible for a UR related to a given drive (such as hunger) to be released by cues compatible with a different drive (such as sex). This

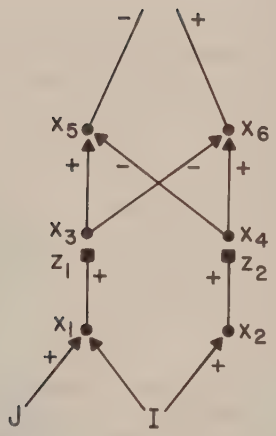

FIG. 24. Network elements that were added to the $\mathcal{A} \rightarrow \mathcal{S}$ part of Figure 23 to produce temporary relief with shock offset. Squares denote accumulation-depletion synapses. Left channel subserved fear; right channel subserved relief. J denotes shock; I denotes tonic arousal. Nonrecurrent inhibition led to output from either x_5 or x_6, not both. (Modified from Grossberg [89]; printed with permission from Elsevier Science Publishing Co., Inc.)

difficulty was overcome by making the $\mathcal{A} \rightarrow \mathcal{S}$ synapses modifiable as well as the $\mathcal{S} \rightarrow \mathcal{A}$ ones. A related difficulty was described by a playful thought experiment: If the network of Figure 23 eats roast turkey for dinner with its lover, it can come away from the dinner table labeling the lover as a source of food and the turkey as a source of sex. Such cross-conditioning could be prevented by combining conditionable $\mathcal{A} \rightarrow \mathcal{S}$ feedback, $U_{12} \rightarrow U_{11}$ feedback that is specific for given sensory representations, and bounds on the total activity of the sensory channels (normalization). As a side effect, these refinements (see Figure 25) allowed the network to model a variety of data on the selective attention of organisms to salient cues, such as the blocking and overshadowing data [117] discussed in Section 6.3.

Since avoidance learning involves inhibitory signals, the selective-attention model carried with it the danger of selective *inattention* to fearful cues. This danger was overcome by making the $\mathcal{A} \rightarrow U_{12}$ signals bifurcate into two signals, one that is excitatory (for promoting attention) and one that can be excitatory or inhibitory (for enhancing or suppressing motor output). This is compatible with recent psychological data [42] showing that a list learned while the subject is *either* in a happy or in a sad mood is recalled best while in the same mood.

Grossberg's article went on to model experimental data on the effects of unexpected versus expected stimuli on the interplay between attention and

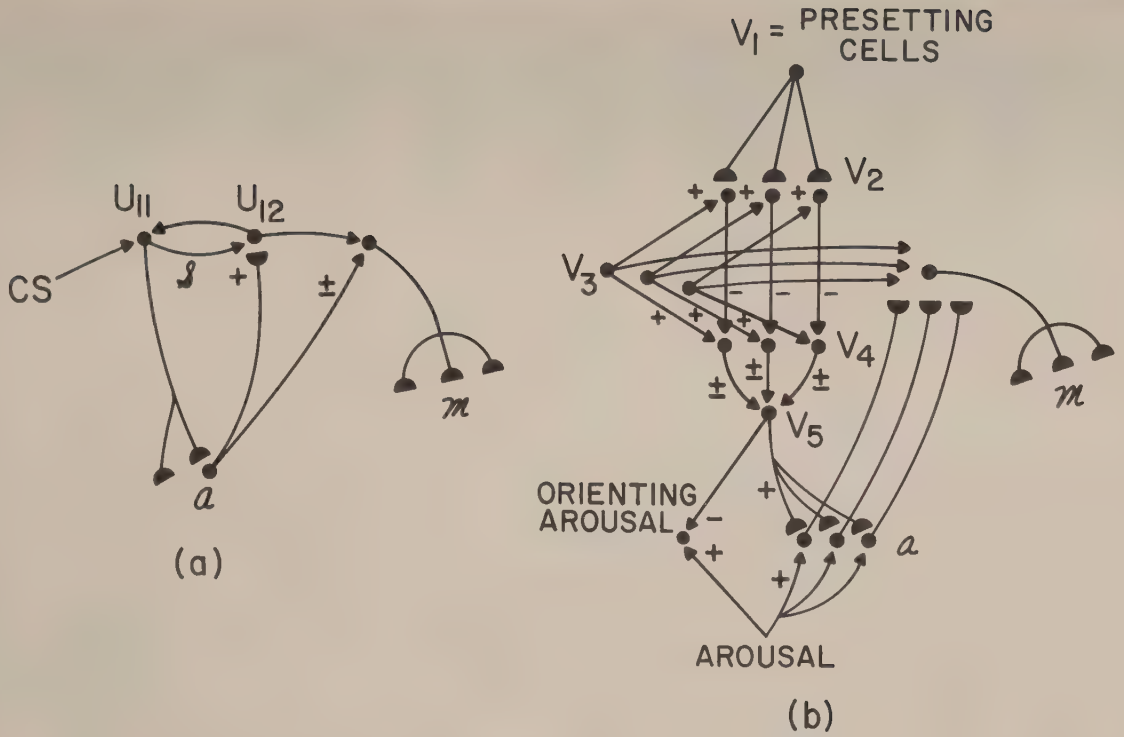

FIG. 25. (a) Network where attentional feedback and motor arousal are in different channels. (From Grossberg [91].) (b) Network where expected events but not unexpected events inhibit orienting arousal. (Modified from Grossberg [91]; printed with permission from Academic Press, Inc.)

arousal. For example, the blocking of CS *X* by CS *A* mentioned above can be overcome by pairing *X* with a different level of shock than *A*. Also, a response-contingent *change* in light intensity, being a novel stimulus, can be used to reward the bar-pressing of a rat [34]. These data led to a question: How can nonspecific arousal be triggered by unexpected but not expected events? One possible answer is shown in Figure 25. A mechanism for learned expectation [87] was discussed in Section 5.2. In Figure 25, a cell responding to a particular expected pattern inhibits nonspecific arousal. Grossberg drew an analogy between the expectation area and the cerebellum, and between the arousal source and the reticular formation. The network appears to be compatible with the suggestion of [162] (see Section 5.2) that the hippocampus is involved in exploratory behavior (as a consequence of unexpected events) and formation of cognitive motor maps.

Other workers, going from 1972 to the present, proposed various forms of a different theory of conditioning, a theory that relied less on detailed anatomies than Grossberg's theory and more on complex processing at the element level. These workers included Klopf [125–127], Rescorla and Wagner [182], Frey and Sears [65], Pearce and Hall [167], Barto and Sutton [31, 32], Barto et al. [33], and Sutton and Barto [199].

Klopf [125] based his theory on a principle which he called *heterostasis*, defined as the opposite of homeostasis. Instead of seeking to keep a certain variable, such as blood sugar or sex hormone, within bounds, the organism sought to *maximize* the value of some variable (here called "pleasure," with "pain" subtracted from it). In his view, not only do brains operate so as to maximize pleasure and minimize pain but so do individual neurons, if "pleasure" is taken to mean depolarization of the cell membrane and "pain" to mean hyperpolarization of the cell membrane.

Operationally, what Klopf called "maximization" was actually a metaphor for a *tendency* in a given direction. This tendency was manifested at the neuronal level by synaptic plasticity of the "heterostatic" type discussed in Section 6.4. Depolarization was considered as "positive reinforcement" for the activity of a particular synapse.

Figure 26 shows how heterostatic synapses fit into Klopf's global brain model. Neurons in all areas sought to maximize received excitation and minimize received inhibition, but the reticular formation and midline thalamus (MTRF) was the final "arbiter" of all decisions. The MTRF coordinated the network's behavior toward the goal of maximizing positive stimulation (pleasure). Since survival is a precondition for pleasure, homeostasis was a major subgoal.

Two other important subsystems in Klopf's network were the limbic system and hypothalamus (LSH) and the neocortex. The LSH sent positive or negative reinforcement to the MTRF, depending on whether or not the MTRF's decisions were favorable to homeostasis. When the MTRF was

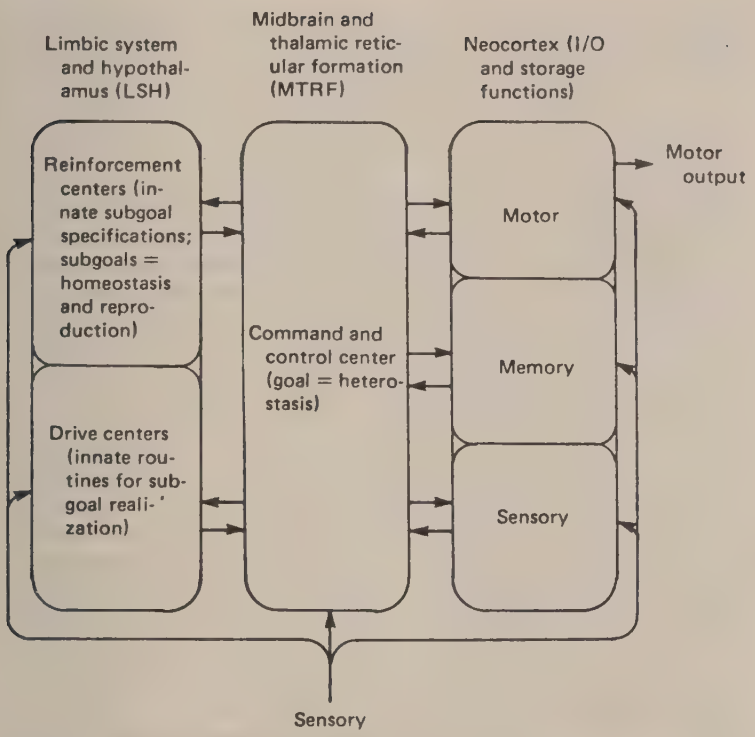
Limbic system and hypothalamus (LSH)
Midbrain and thalamic reticular formation (MTRF)
Neocortex (I/O and storage functions)
Reinforcement centers (innate subgoal specifications; subgoals = homeostasis and reproduction)
Drive centers (innate routines for subgoal realization)
Command and control center (goal = heterostasis)
Motor
Memory
Sensory
Motor output
Sensory

FIG. 26. Model for global organization of brain. (From Klopf [127]; reprinted with permission from Hemisphere Publishing Corporation.)

aroused by stimuli, it sent signals to the neocortex, after which the cortex sent signals back to the MTRF denoting sensory input or memory. Conditioning took place by reinforcement of the cortex so that eventually it could fire without MTRF signals. Selective facilitation of cortical areas corresponded to attention.

Details of Klopf's conditioning mechanism were:

> The sets of neurons that fire in response to the presentation of a CS or US will be denoted by ⟨CS⟩ and ⟨US⟩, respectively. The neuronal model predicts that when these sets are sequentially active within a few seconds of one another, those neurons belonging to the intersection ⟨CS⟩∩⟨US⟩ will be positively reinforced.... Furthermore, the LSH's response to the US is a source of positive reinforcement for neurons in ⟨CS⟩. The effects of both kinds of reinforcement during conditioning will be to cause the sets ⟨CS⟩ and ⟨CS⟩∩⟨US⟩ to increase in size and average frequency of firing. As these increases occur for the set, ⟨CS⟩∩⟨US⟩, one would expect an increasing resemblance between the CR and UR, as indeed happens. [125, p. 29]

Klopf cited a variety of data in support of "heterostatic" synapses. For example, after surface-positive polarization of the motor cortex, a limb can respond to previously ineffective stimuli for up to thirty minutes after polarization ceases [152, 185]. This result does not, however, distinguish a heterostatic rule from an associative rule (cf. Sections 6.1 and 6.2) for synaptic plasticity.

Heterostasis implies positive feedback. Klopf, however, does not distinguish between the normal positive feedback involved in reinforcement and the pathological positive feedback involved in epilepsy. Such a distinction is hard to quantify without terms for the decay of potential and the decay of synaptic efficacy. Both decay terms were absent from Klopf's model.

The heterostatic theory is in line with current trends in the psychology of reinforcement. In 1954, Olds [163] discovered "pleasure" areas in the rat's brain and found that the rat's drive to press a bar for "pleasure"-area stimulation did not satiate with drive satisfaction. The hypothesis that reinforcement is equivalent to drive reduction (for instance, eating makes an animal less hungry) was thus replaced by the hypothesis that "pleasure" areas are part of a positive feedback loop involved in a consummatory response [148]. The levels of body variables, such as blood sugar or sex hormone, are outside the loop but influence it. Thus, electrical stimulation of the lateral hypothalamus makes an animal "hungrier" [50] but is also rewarding [164]. Heterostatic neurons are not the only possible way to explain such phenomena; one alternative explanation [91] was based on positive feedback loops between sensory and drive loci.

Moreover, much conditioning data appears incompatible with the maximization of a single variable:

(1) If an animal first learns a response to escape shock and then is shocked for making that response, self-punitive behavior can result [194].

(2) An animal often prefers intermittent to continuous reinforcement [70]. (This effect is related to human gambling.)

(3) A pigeon that has been rewarded for responding to light of intensity λ_1 and punished for responding to light of intensity $\lambda_2 < \lambda_1$ responds most to intensity $\lambda_3 > \lambda_1$, not at λ_1 [105].

Motor decisions are based on a struggle between the claims of current reinforcement, habits based on past reinforcement, and instinctive patterns. In an effort to explain such complexity, Klopf [127] expanded his model to include "inverse" neurons that "seek" hyperpolarization instead of depolarization.

It was noted in Section 6.4 that Klopf's idea of synaptic change was related to the conditioning model in Rescorla and Wagner [182]. Sutton and

Barto [199] argued that synaptic plasticity of the Klopf or Rescorla-Wagner type could explain two sets of conditioning data better than synaptic plasticity of the associative type, even if time delays were included. One set of data involved the blocking paradigm discussed in Section 6.3. Another set of data involved the observation that the CR often precedes the US. This, they said, was impossible with associative synapses because if a cross-correlation is made between a US and *delayed* CS (which is needed to model the optimal time lag between CS and US), the delay in the CS will also delay the CR by an amount equal to the interstimulus interval. The separate eligibility traces x_i of (6.17) enabled the Sutton-Barto network to overcome this difficulty. A later article [32] is devoted entirely to this phenomenon.

Grossberg's networks with associative synapses can, however, also explain the above conditioning data. Crucial to his approach is the structure of the entire network, not just individual synapses. His explanation of blocking [91] was discussed earlier in this section. Referring to Figure 25, the CR can precede the US for the following reason. Because of the modifiable $U_{11} \to \mathcal{Q}$ and $\mathcal{Q} \to U_{12}$ synapses, repeated CS presentation (activating U_{11}) followed by US presentation (activating $\mathcal{Q}$) causes a buildup of subthreshold activity at U_{12}. Eventually, this activity will be sufficient for the CS to generate the CR along the pathway $U_{11} \to U_{12} \to \mathcal{R}$. The optimal interstimulus interval is determined by the time needed to fixate the CS in short-term memory, which is usually longer than the pathway latencies.

Grossberg [101] gave a critique of the Rescorla-Wagner and related models. In particular, he challenged the statements that "organisms only learn when events violate their expectations" [182, p. 75] and that "a stimulus is likely to be processed to the extent that it is not an accurate predictor of its consequences" [167, p. 538]. He remarked that a US itself—or a CS that has become a US—has expected consequences but is nonetheless processed. He showed that models based on (6.16) could not reconcile the processing of events with unexpected consequences and that of events with expected consequences. In [91], he reconciled them by proposing two separate and competing subsystems—the orienting system and the attention system—for processing the two types of events.

Sutton and Barto stated, in reference to problems of stability and saturation: "Some of the current approaches for solving these problems (notably Grossberg's...) stress the importance of careful network design and use of inhibitory connections.... Although we feel that network level considerations (i.e., *a priori* structure) are very important, they are strongly influenced by the choice of local learning rules, and here we focus on element level issues" [199, p. 159]. The connection between network design and local learning rules is clear from both Section 6 and this subsection. Yet, while "minimal anatomies" are a useful formulation, anatomies should not be made so minimal

that they perform different functions with the same structures when that is inappropriate.

8.2. *CODE DEVELOPMENT: INTERACTIONS BETWEEN SHORT- AND LONG-TERM MEMORY*

In 1973, Malsburg [139] considered the development of tuning of orientation-sensitive cells in the visual cortex. He used unmodifiable recurrent additive excitation and inhibition between cortical cells, and modifiable synapses from a presynaptic (retinal) layer of cells. The equations he used (which had been used for a different purpose in [87]) were

$$\frac{d}{dt}H_k(t) = -\alpha_k H_k(t) + \sum_{j=1}^{n} p_{jk}\left[H_j(t) - \theta_j\right] + \sum_{i=1}^{m} s_{ik}A_i^*(t), \qquad k = 1,\ldots,N, \tag{8.1}$$

where $H_k(t)$ was the excitatory state of cortical cell k at time t, θ_j was the jth cell threshold, $A_i^*(t)$ was the signal of the ith afferent fiber, α_k was a decay rate, and s_{ik} and p_{jk} were connection strengths. The cells were divided up into excitatory or E cells ($p_{jk} > 0$ if cell j is excitatory) and inhibitory or I cells (where $p_{jk} < 0$), as shown in Figure 27.

Malsburg did a numerical simulation of (8.1) with discrete time steps. The synaptic weights s_{ik} were changed at each time step via the rule $\Delta s_{ik} = hA_i^* \max(H_k - \theta_k)$, and then all the s_{ik} were normalized so that $\Sigma_i s_{ik}$ remained constant for each k. This "principle of constant synaptic strengths" had been used earlier by Rosenblatt [184]. Malsburg explained it as follows: "This was introduced for several reasons. One is stability: a system with only growing excitatory synapses is unstable. A second reason is the requirement of high specificity of the cells: they should become insensitive to all stimuli for which they were not trained" [139, p. 99]. He proposed, however, that "both of these functions can probably also be realized by replacing the principle of constant synaptic strengths by additional training of inhibitory coefficients" [139, p. 99].

The Malsburg simulation of the visual cortex exhibited much behavior that is experimentally observed in the actual visual cortex. For example, orientation tuning curves were heavily influenced by development; in fact, the network was insensitive to stimuli not seen during development (cf. [40, 110]). Also, cells sensitive to the same or similar orientation tended to cluster together (see Figure 28). Willshaw and Malsburg [211] extended this theory with computer simulations of square grids of cells at two levels obeying equations similar to (8.1).

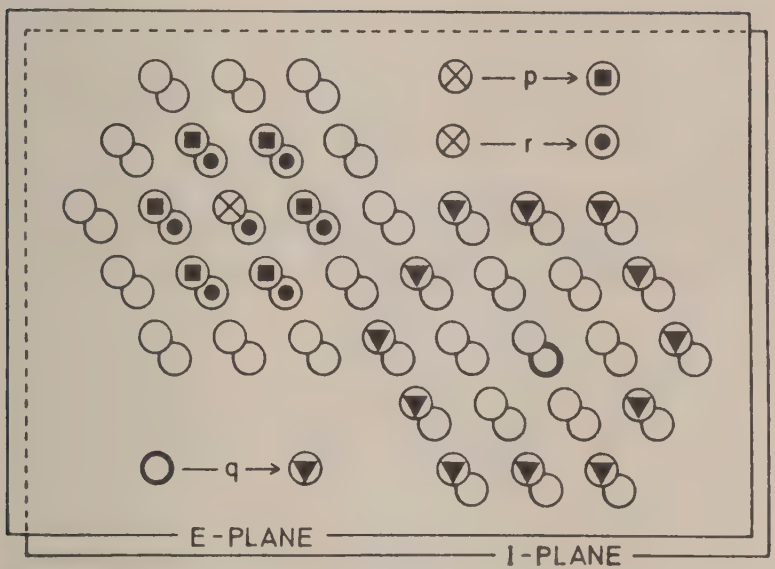

FIG. 27. Small part of simulated cortex. Cells were arranged in hexagonal arrays of E cells (upper plane) and I cells (lower plane). Symbols $\times$, ■, ●, ▼, ○ denote cells which were connected, with connection strengths p, r, and q. (From Malsburg [139]; reprinted with permission from Springer–Verlag.)

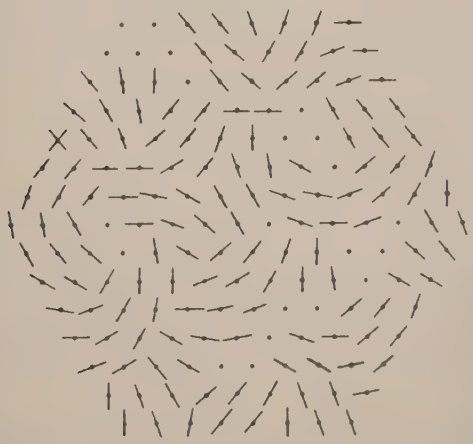

FIG. 28. Preferred orientations of simulated cortical cells after 100 steps of learning. (From Malsburg [139]; reprinted with permission from Springer–Verlag.)

The principle of synaptic conservation was also used in a model of orientation detectors by Perez et al. [171]. (This model contained an on-center off-surround network, not within the cortex but in the retinocortical inputs.) The same principle was used by Wilson [212], for different reasons, in his model of visual spatial frequency adaptation. Here, however, the inhibitory synapses were modifiable and not the excitatory ones. A possible biochemical basis for synaptic conservation was given, namely, that the total quantity of some substance bound to inhibitory transmitter sites should be constant over time. Individual cortical cells were tuned for specific spatial frequencies, and the retina was connected to the cortex by feed-forward excitation and inhibition. The equations were integral equations fairly similar to those used in the early part of [214]. The experimental data modeled included that "threshold elevation curves following adaptation to gratings of high frequency are narrower than those following adaptation to lower spatial frequencies" [212, p. 328], and that "prolonging the period of adaptation profoundly alters the time course of recovery" [212, p. 328].

The same problems of code development and STM-LTM interaction were considered by Grossberg [92–94], in models which built on his own previous work discussed in Sections 6 and 7. The model of [93] used two interacting on-center off-surround fields of cell populations v_1 and v_2 (see Figure 29), which could be interpreted in many possible ways. Not only (as in [139]) could v_1 be retina (or thalamus) and v_2 visual cortex, but also v_1 could be olfactory bulb and v_2 prepyriform cortex; v_1 could be a composite of outer retinal cell layers and v_2 a composite of inner retinal layers; finally, v_1 and v_2 could be two layers within the visual cortex (simple and complex cells, or complex and hypercomplex cells). Thus, the brain could have many layers of such fields, performing ever more intricate feature detection.

The equations in [93] were a synthesis of (6.7) and (7.6). Grossberg let x_{1j} denote the activity of population v_{1j}, x_{2j} the activity of v_{2j}, z_{ij} the synaptic efficacy between v_{1j} and v_{2j}, and $S_j(t)$ the total signal to v_{2j} caused by the normalized spatial pattern $\theta = (\theta_1, \ldots, \theta_n)$ at v_1 (thus $S_j(t) = \sum_{k=1}^{n} \theta_k z_{kj}$). He described the network as follows:

$$\dot{z}_{ij} = (-z_{ij} + \theta_j) x_{2j}, \tag{8.2a}$$

$$x_{2j} = \begin{cases} f(S_j)\left[\sum_{S_k < \varepsilon} f(S_k)\right]^{-1} & \text{if } S_j > \varepsilon, \\ 0 & \text{if } S_j < \varepsilon, \end{cases} \tag{8.2b}$$

where f was an input-output function and ε was a quenching threshold for v_2 (cf. Section 7.2). Equation (8.2b) corresponded to contrast enhancement and noise suppression within the populations v_2.

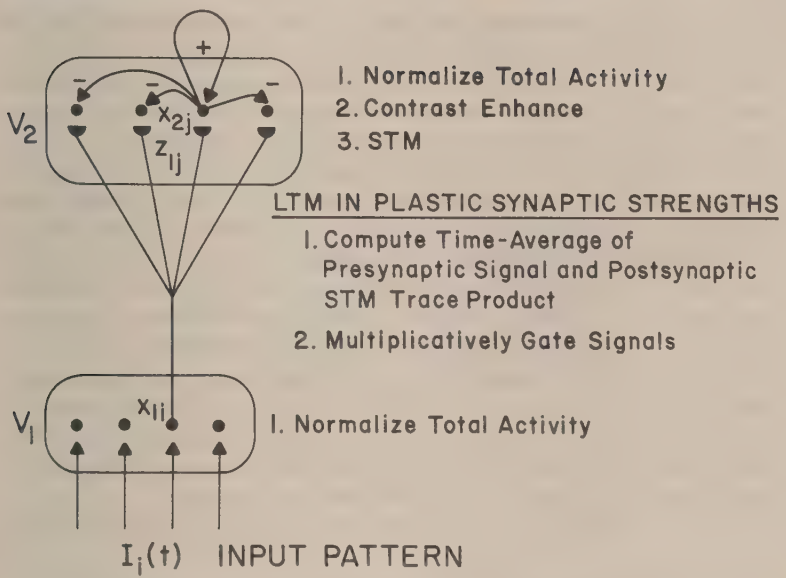

FIG. 29. Network for developmental tuning using two interacting fields with modifiable synapses. Recurrent lateral inhibition takes place within second field. (Modified from Grossberg [93]; reprinted with permission from Springer–Verlag.)

Grossberg analyzed (8.2) under the simplifying condition that only one population stored features in STM, reducing (8.2b) to

$$x_{2j} = \begin{cases} 1 & \text{if} \quad S_j > \max\{\varepsilon, S_k : k \neq j\}, \\ 0 & \text{if} \quad S_j < \max\{\varepsilon, S_k : k \neq j\}. \end{cases}$$

Under that condition, he showed that if j was the unique value such that $x_{2j} = 1$ and the input pattern θ was shown in infinitely many nonoverlapping time intervals, the vector of synaptic weights $\mathbf{z}^{(j)} = (z_{1j}, \ldots, z_{nj})$ converged to θ with ever decreasing angle and $S_j(t)$ was monotonically attracted to $\|\theta\|^2$. Thus, the synaptic weights impinging on j become a code for the pattern θ. Later results in [93] and in [94] gave techniques for coding more than one pattern at a time and for stabilizing such a code in the face of a barrage of other patterns.

Grossberg [92] used the same general principle in discussing the development of cortical feature detectors. He also compared his network with those of Wilson [212] and Malsburg [139]. Grossberg showed that the field effects

of (7.6), with no conditioning but combined with gated dipoles (see Section 8.1), could model the same data that the other two authors had modeled using synaptic conservation and conditionable synapses. He argued, further, that not only was synaptic conservation unnecessary; it could be incompatible with classical conditioning. If a stimulus-response association of the form $S_1 \to R$ was strengthened, and the total weight of all synapses impinging on R was conserved, then some association $S_2 \to R$ with a different stimulus must be weakened. No clear evidence exists that such counterconditioning occurs.

Amari [9] modeled STM-LTM interactions via the interactions between two neural fields, each continuous in one spatial direction with modifiable synapses from one field to another. The presynaptic field was assumed to receive sensory input, and interactions within that field were ignored. The postsynaptic field was assumed to be undergoing distance-dependent recurrent additive lateral interactions which did not change with time, as shown in Figure 30. (These lateral interactions were modeled in a previous article [8].) Amari's model was similar to that of Malsburg [139] except that synaptic conservation was replaced by modifiable inhibitory synapses.

Amari defined these inter- and intrafield interactions using the equations

$$\begin{aligned}
\tau \frac{\partial u(x,t)}{\partial t} &= -u(x,t) + \int w(x-x') f[u(x',t)]\, dx' \\
&\quad + \int s(x,y,t) a(y,t)\, dy - s_0(x,t) - h, \\
\tau \frac{\partial}{\partial t}(x,y,t) &= -s(x,y,t) + ca(y,t)\phi[u(x,t)], \qquad (8.3) \\
\tau'' \frac{\partial}{\partial t} s_0(x,t) &= -s_0(x,t) + c'\phi[u(x,t)],
\end{aligned}$$

where $\phi[u]$ denoted the unit step function, x and y were the spatial coordinates of the post- and presynaptic fields, respectively, $s(x,y,t)$ was the interfield synaptic intensity, $-s_0(x,t)$ was the intensity of a modifiable inhibitory synapse of a constant inhibitory input of intensity 1, f was a bounded monotonic input-output function, $-h$ was the resting potential, $w(x) = w(-x)$ was the recurrent interaction between two cells whose distance apart was x, and $a(y,t)$ was the activity of neurons in the presynaptic field. The function w was of on-center off-surround type: $w(x)$ had its maximum at $x = 0$, then decreased to a negative minimum as x increased, finally approaching 0 as x became large.

Amari studied (8.3) with some realistic mathematical simplifications, namely, that $a(y,t)$ was constant over time t for a given y, τ was small compared to τ' and τ'', and the terms in $a(y,t)\phi[u(x,t)]$ and $\phi[u(x,t)]$ were replaced by their ensemble averages (yielding sigmoid functions). He

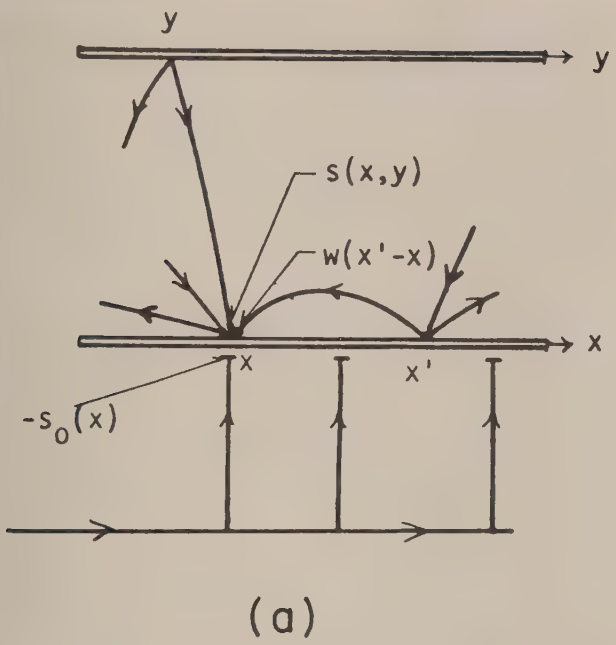

FIG. 30. Model for topographic representation of inputs: (a) schematic of connections between two fields; (b) graph of function w used in part (a). (From Amari [9]; reprinted with permission from Pergamon Press, Ltd.)

proved that a stable equilibrium solution existed of a form that corresponded in neural terms to a topographic map of the input field. The location of points x on the topographic map could be obtained as a function of y by integrating a product of two functions, one dependent on the frequency of stimulation at y, and the other on the intensity distribution of the stimulus.

Linear probabilistic models bearing on feature detection and coding in the visual cortex were proposed in Nass and Cooper [158] and Anderson et al. [20]. These models were based on the linear associative models of [14–17] discussed in Section 6.1.

In [158], a linear mapping existed from one cell collection (corresponding to the retina or the thalamus) to another cell collection (corresponding to the visual cortex). The matrix of synaptic weights defining the mapping changed with correlated pre- and postsynaptic activities. If M was the connectivity matrix and $\mathbf{d}^\beta$ the vector of retinal or geniculate cell activities corresponding to a particular pattern e^β, the ith cell was said to be *committed* to that pattern if the ith component of $M\mathbf{d}^\beta$ was above a certain threshold. In other words, the cell became, from experience, a feature detector for that pattern. Lateral inhibition was introduced to prevent other cells from becoming committed to the same pattern.

In [20], the eigenvectors of the connectivity matrix were considered as equilibrium states of the system and corresponded to distinctive features. In these eigenvectors, activity at each neuron was found to be either its maximum value (taken as 1) or its minimum value (taken as -1). Since the possible activity vectors formed an n-dimensional cube, the model was called the "brain-state-in-a-box model." This model was used to simulate two sets of psychological data. One of these data sets involved perception of linguistic categories, such as discriminating one vowel from another when the sound may be ambiguous. Such categorical perception has been a major concern of artificial-intelligence theorists and is still being studied by Anderson and his coworkers [18]. The other data set involved probability learning, such as learning to predict A eighty percent of the time and B twenty percent of the time after one has heard those letters presented at that frequency. The model simulated probability learning by starting at a random point in n-dimensional space and calculating the probabilities of converging to different corners of the box.

In both [158] and [20] activity of each neuron was arbitrarily bounded by fixing activity when it reached a given value. Such "hard saturation" was necessary to prevent instability, because the system was linear with positive eigenvalues. Grossberg [98] showed that hard saturation, as opposed to normalization, caused inherent difficulties for pattern processing. In particular, relative weights in input patterns could be distorted; in other words, the asymptotic short-term storage of the more significant parts of a pattern that occurred in [90] (see Section 7.2) could not be achieved in the model of

Anderson et al. Anderson and Silverstein [19] conceded that the model of [20] had such limitations, but argued that linear transduction at many levels of sensory systems has been experimentally verified. Moreover, they argued that the model of [20] indicated some interesting phenomena that were likely to remain true in less simplified models.

It is true that the "understanding of brain functions poses problems of sufficient magnitude to allow a variety of useful approaches" [19, p. 602]. Yet, it is now possible to study models that synthesize linear and nonlinear effects. Recall from Section 7.1, for example, that the linear Hartline-Ratliff [108] equations arose from the steady-state behavior of nonlinear equations due to Grossberg [78]. It is an unfinished, and useful, task to simulate such phenomena as category perception using models that also include specific nonlinearities suggested by biological data. (Another nonlinear model of category perception appears in [12].)

Bienenstock et al. [38] introduced nonlinear effects into a feature-detection model that was otherwise along the lines of [158]. Their model included a random input $(d_1(t),\ldots,d_n(t))$ (from the retinal level), a cortical neuronal response $c(t)=\Sigma_j m_j(t)d_j(t)$, and synaptic weights changing according to the rule $\dot{m}_j(t)=-\varepsilon m_j(t)+\phi(c(t), E(c(t))d(t)$, where ϕ was any of a class of nonlinear functions which equal zero for c below some threshold, and E denoted expected value. They showed that if different input patterns came in that were nearly orthogonal in n-dimensional space, lateral inhibitory connections between neurons could be introduced so as to make each neuron selectively responsive to a different pattern.

The stochastic model of [38] was used to simulate the effects of early visual experience on adult binocular vision. Many of the phenomena found were also found in simulations by Cohen and Grossberg [47] of a deterministic model based on [92–94].

The topographical connection of two successive sensory layers has also been studied at the level of the frog retina and tectum (for example, [28, 133, 166]), where the anatomy is better understood than in the mammalian neocortex. The frog tectum is homologous in structure to the mammalian superior colliculus, a secondary visual processing area, but its function is more that of a primary visual processing area. Thus, retinotectal map studies contribute to the understanding of possible retinocortical maps.

8.3. *GENERAL REMARKS ON MULTILEVEL BRAIN MODELS*

The work discussed in this chapter suggests that substantial progress has been made toward "global brain models" that are plausible and predictive of data. Brain areas like the reticular formation, hypothalamus, and limbic system—though harder to understand mathematically than the sensory and motor areas—have been embedded in multilevel networks with a precise

qualitative description. The real-time behavior of such networks in response to various sequences of stimuli is now undergoing computer simulations. Network studies have begun to suggest interdisciplinary experiments to resolve disputes on the behavioral functions of these brain areas.

Such networks are currently being extended to include models of complex goal-directed behavioral sequences. Goal-directed activity, which is a major concern of artificial-intelligence theorists, is also being widely studied by neural theorists [3, 26, 96, 129, 189] and neural experimentalists [66, 156].

Many scientists have proposed qualitative frameworks for the global organization of behavior. Powers [174], for example, explained behavior as a hierarchy of levels of processing: intensity, sensations, configurations, transitions, sequences, relationships, programs, system concepts. The time is perhaps ripe for quantifying such frameworks by means of other global frameworks built on the dynamics of detailed anatomies.

9. CONCLUDING REMARKS: WHY LARGE-SCALE NEURAL MODELS ARE USEFUL

The results discussed in the first eight sections show that neural models have imposed some order on the vast panoply of vertebrate psychological functions. Arbib's prediction in 1975 that "perhaps...five years from now one can confidently speak of 'brain theory' as a well-defined field with an accepted core of basic knowledge" [25, p. 280] has partially come to pass. Details of the "core of basic knowledge" are still in dispute, but modelers have reached a consensus on many of the basic principles and thus have come much closer to forming a coherent international community. This trend is also indicated by the history of the *Brain Theory Newsletter*, which was initiated in 1975 to stimulate dialogue in the field. By 1981, it had expanded to a full-fledged journal, *Brain Theory and Cognition*. This journal and its predecessor have counteracted the occasional tendency for researchers in the field not to address valuable work that bears on the same problems as their own work. Another journal devoted to this field, the *Journal of Theoretical Neurobiology*, began publication in 1981. Articles on neural modeling also appear frequently in journals with a wider audience, notably *Biological Cybernetics*, *Mathematical Biosciences*, and the *Journal of Theoretical Biology*.

The partnership between neural theorists and experimentalists has grown steadily over the last several years. Strengthening such partnership is the joint responsibility of both groups. Whereas theorists have at times neglected experimental data, experimentalists have at times resisted general theory. Part of the resistance to theory is an overreaction to an earlier overenthusiasm. In the 1940s and 1950s, many experimental biologists expected too much, perhaps, of the new cybernetic formulations then arising. When their expectations were unfulfilled, some lost faith in the utility of neural models

altogether, with the exception of single-cell models. The historical outline presented in this article argues that this loss of faith was unwarranted. Neural theory in forty years, while building on the work of its early pioneers, has gone far beyond them. Thus, contemporary theoretical work should be judged on its own merits, not by the limitations of the early work.

Another part of the resistance to theory comes from a respect for the great diversity of neural functions in many phyla of animals (see, for example, [172]). Also, biologists often advise that the mathematician should first try to model invertebrate nervous systems rather than vertebrate nervous systems because invertebrate systems are "simpler."

The "diversity" argument was answered by Grossberg as follows: "You cannot have variations on a theme unless you have a theme" (personal communication). The central task of the modeler is thus the search for themes relevant to the flexible adaptation of an organism's responses to a complex and changing environment.

The "start with invertebrates" argument was answered by Arbib et al. as follows:

> Vertebrate strategies of learning by the *organism* may be *essentially* different from those of invertebrates. Thus studies of learning in such invertebrate preparations as *Aplysia* may only contribute to mammalian studies at the level of determining the mechanisms of neural modification, rather than at the level of seeing how such modifications may contribute to the change in network behavior. [27, p. 111; authors' italics]

At the single-cell and small-network levels, classical invertebrate studies have contributed vital insights that remain valid for mammals. Examples are the studies of the squid giant axon [121] and of the horseshoe-crab eye [108]. At the global level, however, the themes of invertebrate behavior are likely to be far from our intuition, which is, of course, derived from human introspection. Moreover, invertebrate single neurons and small networks appear to be as complex as those of vertebrates.

In the search for themes, distinctions between different kinds of neural models become crucial. If, say, *linear* or *digital-computer* models of a particular phenomenon have proved inadequate, it is wrong to conclude that *all* mathematical models of the phenomenon will be inadequate (cf. the discussion in Section 8.2 on linear versus nonlinear models of visual processing).

Biologists are most interested in a theory when it is immediately testable experimentally. The concern for testability is essential for good modeling, but as many examples from physics show, there can be a time lag of years between development of a theory and its experimental proof or disproof. Thus, a theory can be a contribution to knowledge even if some of its variables are hard to localize in the brain with current experimental techniques. In fact, many testable predictions have already been made by neural theorists, but experimental psychologists or physiologists are not always

equipped to test them. Some experimentalists do not understand the distinctions between competing theories; others are eager to test a theory but are forced by current research economics to do "safe" experiments with short turnover time instead. The comfortable interplay that exists in physics between experiment and theory, each with a life of its own, has yet to be established in neurobiology. Theoretical biology is, however, evolving toward autonomous existence as a subdiscipline in much the same manner as theoretical physics.

Experiments likely to yield insights about the validity of given population models are often hard to perform because they have a long turnover time and require coordination of large amounts of data. Yet, a growing number of neurobiologists are interested enough in the basic questions to risk doing such work. For example, Freeman [64] has done parallel electrode recording over large areas of the neocortex. His data relate certain firing patterns to attention and expectation, for example, the synchronization of cortical activity caused by smelling an expected scent. Donchin and his coworkers [53, 196] have extensively studied two competing EEG patterns related to expectancy: the contingent negative variation and P300. Mountcastle [155] has found that selective lesions of the posterior parietal lobe in monkeys and humans can disrupt motor commands. Other examples are too numerous to mention here; some are discussed elsewhere in this article.

Such advances in gross potential recording and in physiological psychology bode well for future interactions between theory and experiment. These interactions will also be stimulated by current exciting research in neuronal chemistry. Examples include the cyclic-nucleotide work mentioned in Section 6, and work on the role of endorphins and small polypeptides in neural transmission at various brain loci [41].

These results and many others need to be fitted into a unified conceptual framework, a framework soundly rooted in observation, logic, and mathematical reasoning. Neural population modeling provides the beginnings of such a framework. As it matures, it will yield both insights into age-old philosophical questions about the mind and applications of great practical importance. As part of a general increase in the understanding of brain function, it will provide long-term contributions to neurological and psychiatric treatment. For example, networks have been developed in which symptoms of schizophrenia can be obtained from general overarousal and symptoms of Parkinson's disease from underarousal [88, 89, 99, 104]. Moreover, the effects of some drugs on each condition in these networks are the same as in humans. The updated neural modeling of the last decade can also, perhaps, contribute in the short term to certain industrial applications of artificial intelligence (cf. Section 3.2). If the false hopes that arose during the early "computer-brain" days can be avoided, the time is ripe for a renewed partnership between brain theory and nonbiological adaptive-system theory.

Neural population modeling essentially began in the cybernetic heyday of Wiener, von Neumann, McCulloch, and Rashevsky, and has evolved and grown in the intervening years. Continued progress depends on the joint effort of "systems-oriented" researchers in many disciplines, including neurobiology, mathematics, experimental psychology, clinical neurology, psychiatry, electrical engineering, and computer science. It is essential that all these workers constantly address and criticize each other's ideas. I hope that this review article will contribute to the spirit of cooperation and friendly competition within the network of neural modelers.

I wish to thank the following persons who have helped make this work possible. Stephen Grossberg gave valuable comments at all stages of the manuscript, and helpful suggestions for the overall direction of the article. James Anderson, Michael Arbib, Andrew Barto, Stuart Geman, and Harry Klopf provided constructive criticism and clarified the import of their own research. David Dahlbacka did a precise editing of the manuscript which markedly improved the exposition. Virginia Perrenod typed several versions of the manuscript superbly, including the final draft. Original illustrations and redrawings were done by Moon Vanko, and reproductions were done by the Audiovisual Services at the University of Houston. Financial support for these services, and for travel to consult with colleagues, was provided by a Limited Grant-in-Aid from the Office of Sponsored Programs, University of Houston. Lorraine Levine provided needed humor and moral support.

REFERENCES

1 E. D. Adrian, *The Mechanism of Nervous Action*, Oxford U.P., London, 1932.
2 J. S. Albus, A theory of cerebellar function, *Math. Biosci.* 10:25–61 (1971).
3 J. S. Albus, Mechanisms of planning and problem solving in the brain, *Math. Biosci.* 45:247–293 (1979).
4 S. Amari, Characteristics of randomly connected threshold element networks and network systems, *Proc. IEEE* 59:35–47 (1971).
5 S. Amari, Characteristics of random nets of analog neuron-like elements, *IEEE Trans. Systems Man Cybernet.* SMC-2:563–657 (1972).
6 S. Amari, A method of statistical neurodynamics, *Kybernetik* 14:201–215 (1974).
7 S. Amari, Neural theory of association and concept-formation, *Biol. Cybernet.* 26:175–185 (1977).
8 S. Amari, Dynamics of pattern formation in lateral-inhibition type neural fields, *Biol. Cybernet.* 27:77–87 (1977).
9 S. Amari, Topographic organization of nerve fields, *Bull. Math. Biol.* 42:339–364 (1980).
10 S. Amari and M. A. Arbib, Competition and cooperation in neural nets, in *Systems Neuroscience* (J. Metzler, Ed.), Academic, New York, 1977, pp. 119–165.
11 S. Amari and M. A. Arbib (Eds.), *Competition and Cooperation in Neural Nets*, Lecture Notes in Biomathematics, Vol. 45, Springer, New York, 1982.
12 S. Amari and M. Takeuchi, Mathematical theory on formation of category detecting nerve cells, *Biol. Cybernet.* 27:127–136 (1978).

13 P. Andersen and J. C. Eccles, Inhibitory phasing of neural discharge, *Nature* 196:645–647 (1962).
14 J. A. Anderson, A memory storage model utilizing spatial correlation functions, *Kybernetik* 5:113–119 (1968).
15 J. A. Anderson, Two models for memory organization using interacting traces, *Math. Biosci.* 8:137–160 (1970).
16 J. A. Anderson, A simple neural network generating an interactive memory, *Math. Biosci.* 14:197–220 (1972).
17 J. A. Anderson, A theory for the recognition of items from short memorized lists, *Psych. Rev.* 80:417–438 (1973).
18 J. A. Anderson, Cognitive and psychological computation with neural models, *IEEE Trans. Systems Man. Cybernet.*, to appear.
19 J. A. Anderson and J. W. Silverstein, Reply to Grossberg, *Psych. Rev.* 85:597–603 (1978).
20 J. A. Anderson, J. W. Silverstein, S. A. Ritz, and R. S. Jones, Distinctive features, categorical perception, and probability learning: Some applications of a neural model, *Psych. Rev.* 84:413–451 (1977).
21 P. A. Anninos, B. Beek, T. J. Csermely, E. M. Harth, and G. Pertile, Dynamics of neural structures, *J. Theoret. Biol.* 26:121–148 (1970).
22 P. A. Anninos, Mathematical models of memory traces and forgetfulness, *Kybernetik* 10:165–167 (1972).
23 P. A. Anninos, Cyclic modes in artificial neural nets, *Kybernetik* 11:5–14 (1972).
24 M. A. Arbib, *The Metaphorical Brain*, Wiley-Interscience, New York, 1972.
25 M. A. Arbib, From automata theory to brain theory, *Internat. J. Man-Mach. Stud.* 7:279–295 (1975).
26 M. A. Arbib, Visuomotor coordination: From neural nets to schema theory, *Cogn. and Brain Theory* 4:23–39 (1980).
27 M. A. Arbib, W. L. Kilmer, and R. N. Spinelli, Neural models and memory, in *Neural Mechanisms and Memory* (M. R. Rosenzweig and E. L. Bennett, Eds.), MIT Press, Cambridge, Mass., 1976.
28 M. A. Arbib and R. Lara, A neural model of the interaction of tectal columns in prey-catching behavior, *Biol. Cybernet.* 44:185–196 (1982).
29 W. R. Ashby, H. von Foerster, and C. C. Walker, Instability of pulse activity in a net with threshold, *Nature* 196:561–562 (1962).
30 D. H. Ballard and C. M. Brown, *Computer Vision*, Prentice-Hall, Englewood Cliffs, N. J., 1982.
31 A. G. Barto and R. S. Sutton, Goal seeking components for adaptive intelligence assessment, Air Force Technical Report AFWAL-TR-81-1070, Dayton, Ohio, 1981.
32 A. G. Barto and R. S. Sutton, Simulation of anticipatory responses in classical conditioning by a neuron-like adaptive element, *Behav. Brain Res.* 4:221–235 (1982).
33 A. G. Barto, R. S. Sutton, and P. Brouwer, Associative search network: A reinforcement learning associative memory, *Biol. Cybernet.* 40:201–211 (1981).
34 D. E. Berlyne, Attention as a problem in behavior theory, in *Attention: Contemporary Theory and Analysis* (D. E. Mostofsky, Ed.), Appleton-Century-Crofts, New York, 1970, p. 25.
35 R. L. Beurle, Properties of a mass of cells capable of regenerating pulses, *Philos. Trans. Roy. Soc. London Ser. B* 250:55–84 (1956).
36 R. L. Beurle, Functional organization in random networks, in *Principles of Self-Organization* (H. von Foerster and G. W. Zopf, Eds.), Pergamon, New York, 1962, p. 291.

37 R. L. Beurle, Information in random networks, in *Aspects of the Theory of Artificial Intelligence* (C. A. Muses, Ed.), Plenum, New York, 1962, p. 19.
38 E. L. Bienenstock, L. N. Cooper, and P. W. Munro, Theory for the development of neuron selectivity: Orientation specificity and binocular interaction in visual cortex, *J. Neurosci.* 2:32–48 (1982).
39 C. Blakemore, R. H. S. Carpenter, and M. A. Georgeson, Lateral inhibition between orientation detectors in the human visual system, *Nature* 228:37–39 (1970).
40 C. Blakemore and G. F. Cooper, Development of the brain depends on the visual environment, *Nature* 228:477–478 (1970).
41 R. C. Bolles and M. S. Fanselow, Endorphins and behavior, *Ann. Rev. Psych.* 33:87–101 (1982).
42 G. H. Bower, K. P. Monteiro, and S. G. Gilligan, Emotional mood as a context for learning and recall, *J. Verbal Learn. and Verbal Behav.* 17:573–585 (1978).
43 C. C. Boylls, A theory of cerebellar function with applications to locomotion. I. The physiological role of climbing fiber inputs in anterior lobe operation, Technical Report 75C-6, Computer and Information Science Dept., Univ. of Massachusetts at Amherst, 1975.
44 G. S Brindley, The classification of modifiable synapses and their use in models of conditioning, *Proc. Roy. Soc. London Ser. B* 168:361–376 (1967).
45 G. S. Brindley, Nerve net models of plausible size that perform many simple learning tasks, *Proc. Roy. Soc. London Ser. B* 174:173–191 (1969).
46 G. G. Carpenter, A geometric approach to singular perturbation problems with applications to nerve impulse equations, *J. Differential Equations* 23:335–367 (1977).
47 M. A. Cohen and S. Grossberg, Some global properties of binocular resonances: Disparity scaling, filling-in and figure-ground, in *Figural Synthesis* (T. Caelli and P. Dodwell, Eds.), Erlbaum, Hillside, N.J., 1982.
48 M. A. Cohen and S. Grossberg, Absolute stability of global pattern formation and parallel memory storage by competitive neural networks, submitted to *IEEE Trans. Systems Man and Cybernet.*
49 J. D. Cowan, A statistical mechanics of nervous activity, in *Lectures on Mathematics in the Life Sciences* (M. Gerstenhaber, Ed.), Vol. 2, Amer. Math. Soc., Providence, R.I., 1970.
50 J. M. R. Delgado and B. K. Anand, Increase of food intake induced by electrical stimulation of the lateral hypothalamus, *Amer. J. Physiol.* 172:162–168 (1953).
51 P. Dev, Perception of depth surfaces in random-dot stereograms: A neural model, *Internat. J. Man-Mach. Stud.* 7:511–528 (1975).
52 R. L. Didday, The simulation and modeling of distributed information processing in the frog visual system, Ph.D. Thesis, Stanford Univ., 1970.
53 E. Donchin, P. Tueting, W. Ritter, M. Kutas, and E. Heffley, On the independence of the CNV and P300 components of the human average evoked potential, *EEG Clin. Neurophys.* 38:449–462 (1975).
54 J. C. Eccles, An instruction-selection theory of learning in the cerebellar cortex, *Brain Res.* 127:327–352 (1977).
55 J. C. Eccles, M. Ito, and J. Szentagothai, *The Cerebellum As a Neuronal Machine*, Springer, New York, 1967.
56 S. A. Ellias and S. Grossberg, Pattern formation, contrast control and oscillations in the short-term memory of shunting on-center off-surround networks, *Biol. Cybernet.* 20:69–98 (1975).
57 G. B. Ermentrout and J. D. Cowan, Large scale spatially organized activity in neural nets, *SIAM J. Appl. Math.* 38:1–21 (1980).

58 J. W. Evans, Nerve axon equations, I: Linear approximations, *Indiana Univ. Math. J.* 21:877–885 (1972).
59 D. H. Fender and B. Julesz, Extension of Panum's fusional area in binocularly stabilized vision, *J. Opt. Soc. Amer.* 57:819–830 (1967).
60 R. R. FitzHugh, Impulses and physiological states in theoretical models of nerve membrane, *Biophys. J.* 1:445–466 (1961).
61 W. J. Freeman, Waves, pulses, and the theory of neural masses, *Progr. Theoret. Biol.* 2:87–165 (1972).
62 W. J. Freeman, Linear analysis of the dynamics of neural masses, *Ann. Rev. Biophys. Bioeng.* 1:225–256 (1972).
63 W. J. Freeman, Parallel processing of signals in neural sets as manifested in the EEG, *Internat. J. Man-Mach. Stud.* 7:347–369 (1975).
64 W. J. Freeman, *Mass Action in the Nervous System*, Academic, New York, 1975.
65 P. W. Frey and R. J. Sears, Model of conditioning incorporating the Rescorla-Wagner associative axiom, a dynamic attention rule and a catastrophe rule, *Psych. Rev.* 85:321–340 (1978).
66 J. Fuster, *The Prefrontal Cortex*, Raven Press, New York, 1980.
67 S. Geman, Some averaging and stability results for random differential equations, *SIAM J. Appl. Math.* 36:86–105 (1979).
68 S. Geman, The law of large numbers in neural modeling, in *Mathematical Psychology and Psychophysiology* (S. Grossberg, Ed.), Amer. Math. Soc., Providence, R.I., 1980.
69 J. J. Gibson and M. Radner, Adaptation, after-effect, and contrast in the perception of tilted lines. I. Quantitative studies, *J. Exp. Psych.* 20:453–467 (1937).
70 J. A. Gray and P. T. Smith, in *Animal Discrimination Learning* (R. M. Gilbert and N. S. Sutherland, Eds.), Academic, New York, 1969, p. 243.
71 P. Greengard, Possible role for cyclic nucleotides and phosphorylated membrane proteins in postsynaptic actions of neurotransmitters, *Nature* 260:101–108 (1976).
72 J. S. Griffith, On the stability of brain-like structures, *Biophys. J.* 3:299–308 (1963).
73 J. S. Griffith, A field theory for neural nets: I: Derivation of the field equations, *Bull. Math. Biophys.* 25:111–120 (1963).
74 J. S. Griffith, A field theory for neural nets: II: Properties of the field equations, *Bull. Math. Biophys.* 17:187–195 (1965).
75 S. Grossberg, A prediction theory for some non-linear functional-differential equations: I. Learning of lists, *J. Math. Anal. Appl.* 21:643–694 (1968).
76 S. Grossberg, A prediction theory for some non-linear functional-differential equations: II. Learning of patterns, *J. Math. Anal. Appl.* 22:490–522 (1968).
77 S. Grossberg, Embedding fields: A theory of learning with physiological implications, *J. Math. Psych.* 6:209–239 (1969).
78 S. Grossberg, On learning, information, lateral inhibition, and transmitters, *Math. Biosci.* 4:255–310 (1969).
79 S. Grossberg, Some networks that can learn, remember, and reproduce any number of complicated space-time patterns, I, *J. Math. Mech.* 19:53–91 (1969).
80 S. Grossberg, On the production and release of chemical transmitters and related topics in cellular control, *J. Theoret. Biol.* 22:325–364 (1969).
81 S. Grossberg, On learning and energy-entropy dependence in recurrent and nonrecurrent signed networks, *J. Statist. Phys.* 1:319–350 (1969).
82 S. Grossberg, On the global limits and oscillations of a system of nonlinear differential equations describing a flow on a probabilistic network, *J. Differential Equations* 5:531–563 (1969).

83 S. Grossberg, Some networks that can learn, remember, and reproduce any number of complicated space-time patterns, II. *Stud. Appl. Math.* 49:135–166 (1970).
84 S. Grossberg, Neural pattern discrimination, *J. Theoret. Biol.* 27:291–337 (1970).
85 S. Grossberg, On the dynamics of operant conditioning, *J. Theoret. Biol.* 33:225–255 (1971).
86 S. Grossberg, Pattern learning by functional-differential neural networks with arbitrary path weights, in *Delay and Functional Differential Equations and Their Applications* (K. Schmitt, Ed.), Academic, New York, 1972, pp. 121–160.
87 S. Grossberg, Cerebellar and retinal analogs of cells fired by learnable or unlearned pattern classes, *Kybernetik* 10:49–57 (1972).
88 S. Grossberg, A neural theory of punishment and avoidance, I. Qualitative theory, *Math. Biosci.* 15:39–67 (1972).
89 S. Grossberg, A neural theory of punishment and avoidance, II. Quantitative theory, *Math. Biosci.* 15:253–285 (1972).
90 S. Grossberg, Contour enhancement, short term memory, and constancies in reverberating neural networks, *Stud. Appl. Math.* 52:213–257 (1973).
91 S. Grossberg, A neural model of attention, reinforcement, and discrimination learning, *Internat. Rev. Neurobiol.* 18:263–327 (1975).
92 S. Grossberg, On the development of feature detectors in the visual cortex with applications to learning and reaction-diffusion systems, *Biol. Cybernet.* 21:145–159 (1976).
93 S. Grossberg, Adaptive pattern classification and universal recoding: I. Parallel development and coding of neural feature detectors, *Biol. Cybernet.* 23:121–134 (1976).
94 S. Grossberg, Adaptive pattern classification and universal recoding: II. Feedback, expectation, olfaction, illusions, *Biol. Cybernet.* 23:187–202 (1976).
95 S. Grossberg, Pattern formation by the global limits of a nonlinear competitive interaction in *n* dimensions, *J. Math. Biol.* 4:237–256 (1977).
96 S. Grossberg, A theory of human memory: Self-organization and performance of sensory-motor codes, maps, and plans, *Progr. Theoret. Biol.* 5:233–374 (1978).
97 S. Grossberg, Competition, decision, and consensus, *J. Math. Anal. Appl.* 66:470–493 (1978).
98 S. Grossberg, Do all neural models really look alike? A comment on Anderson, Silverstein, Ritz, and Jones, *Psych. Rev.* 85:592–596 (1978).
99 S. Grossberg, How does the brain build a cognitive code? *Psych. Rev.* 87:1–51 (1980).
100 S. Grossberg, Adaptive resonance in development, perception, and cognition, in *Mathematical Psychology and Psychophysiology* (S. Grossberg, Ed.), Amer. Math. Soc., Providence, R.I., 1980.
101 S. Grossberg, The processing of expected and unexpected events during conditioning and attention: A psychophysiological theory, to appear in *Psych. Rev.*
102 S. Grossberg, The quantized geometry of visual space: The coherent computation of depth, form, and lightness, to appear in *Behav. Brain Sci.*
103 S. Grossberg and D. S. Levine, Some developmental and attentional biases in the contrast enhancement and short-term memory of recurrent neural networks, *J. Theoret. Biol.* 53:341–380 (1975).
104 S. Grossberg and J. Pepe, Schizophrenia: Possible dependence of associational span, bowing, and primacy vs. recency on spiking threshold, *Behav. Sci.* 15:359–362 (1970).
105 H. M. Hanson, Effects of discrimination training on stimulus generalization, *J. Exp. Psych.* 58:321–334 (1969).

106 L. D. Harmon and E. R. Lewis, Neural modeling, *Physiol. Rev.* 46:513–591 (1968).
107 E. M. Harth, T. J. Csermely, B. Beck, and R. D. Lindsay, Brain functions and neural dynamics, *J. Theoret. Biol.* 26:93–120 (1970).
108 H. K. Hartline and F. Ratliff, Inhibitory interaction of receptor units in the eye of *Limulus*, *J. Gen. Physiol.* 40:351 (1957).
109 D. O. Hebb, *The Organization of Behavior*, Wiley, New York, 1949.
110 H. V. B. Hirsch and D. N. Spinelli, Visual experience modifies distribution of horizontally and vertically oriented receptive fields in cats, *Science* 168:869–871 (1970).
111 A. L. Hodgkin, *The Conduction of the Nervous Impulse*, C. C. Thomas, Springfield, Ill., 1964.
112 A. L. Hodgkin and A. F. Huxley, A quantitative description of membrane current and its application to conduction and excitation in nerve, *J. Physiol.* 117:500–544 (1952).
113 D. H. Hubel and T. N. Wiesel, Receptive fields, binocular interaction, and functional architecture in the cat's visual cortex, *J. Physiol.* 160:106–154 (1962).
114 D. H. Hubel and T. N. Wiesel, Receptive fields and functional architecture in two non-striate visual areas (18 and 19) of the cat, *J. Neurophys.* 28:229–298 (1965).
115 C. L. Hull, *Principles of Behavior*, Appleton, New York, 1943.
116 M. Ito, Mechanisms of motor learning, in *Competition and Cooperation in Neural Nets*, Lecture Notes in Biomathematics, Vol. 45 (S. Amari and M. A. Arbib, Eds.), Springer, New York, 1982, pp. 418–429.
117 L. J. Kamin, Predictability, surprise, attention, and conditioning, in *Punishment and Aversive Behavior* (B. A. Campbell and R. M. Church, Eds.), Appleton-Century-Crofts, New York, 1969, p. 279.
118 E. R. Kandel, *A Cell-biological Approach to Learning*, Grass Lecture Monograph 1, Soc. for Neuroscience, Bethesda, Md., 1978.
119 T. Kasamatsu and J. D. Pettigrew, Depletion of brain catecholamines: Failure of ocular dominance shift after monocular occlusion in kittens, *Science* 194:206–209 (1976).
120 A. Katchalsky, V. Rowland, and R. Blumenthal (Eds.), Dynamic patterns of brain cell assemblies, *Neurosci. Res. Progr. Bull.* 12:3–187 (1974).
121 B. Katz, *Nerve, Muscle, and Synapse*, McGraw-Hill, New York, 1966.
122 W. L. Kilmer, W. S. McCulloch, and J. Blum, A model of the vertebrate central command system, *Internat. J. Man-Mach. Stud.* 1:279–309 (1969).
123 W. L. Kilmer and T. McLardy, A diffusely programmed but sharply trainable hippocampus model, *Internat. J. Neurosci.* 2:241–248 (1971).
124 W. L. Kilmer and M. Olinski, Model of a plausible learning scheme for CA3 hippocampus, *Kybernetik* 16:133–143 (1974).
125 A. H. Klopf, Brain function and adaptive systems: A heterostatic theory, Air Force Cambridge Research Laboratories Research Report AFCRL-72-0164, Bedford, Mass., 1972; summary in *Proceedings of the International Conference on Systems, Man, and Cybernetics*, IEEE Syst., Man, and Cybernet. Soc., Dallas, 1974.
126 A. H. Klopf, Goal-seeking systems from goal-seeking components, *Cogn. and Brain Theory. Newsl.* 3, No. 2 (1979).
127 A. H. Klopf, *The Hedonistic Neuron*, Hemisphere, Washington, 1982.
128 L. Koenigsberger, *Hermann von Helmholtz*, translated by F. A. Welby, Dover, New York, 1965.

129 T. Kohonen, *Associative Memory — A System-Theoretical Approach*, Springer, New York, 1977.

130 T. Kohonen, A simple paradigm for the self-organized formation of structured feature maps, in *Competition and Cooperation in Neural Nets*, Lecture Notes in Biomathematics, Vol. 45 (S. Amari and M. A. Arbib, Eds.), Springer, New York, 1982, pp. 248–265.

131 T. Kohonen, P. Lehtiö, J. Rovamo, J. Hyvärinen, K. Bry, and L. Vainio, A principle of neural associative memory, *Neuroscience* 2:1065–1076 (1977).

132 S. Kuffler, Discharge patterns and functional organization of mammalian retina, *J. Neurophys.* 16:37–68 (1953).

133 R. Lara and M. A. Arbib, A neural model of interaction between pretectum and tectum in prey selection, *Cogn. Brain Theory* 5:149–171 (1982).

134 K. Lashley, *Brain Mechanisms and Intelligence*, Univ. of Chicago Press, Chicago, 1929.

135 D. S. Levine, Studies in the transformation and storage of patterns in reverberating neural networks, Ph.D. thesis, Massachusetts Institute of Technology, 1975, pp. 127–136.

136 D. S. Levine, Existence of a limiting pattern for a system of nonlinear equations describing inter-population competition, *Bull. Math. Biol.* 41:617–628 (1979).

137 D. S. Levine and S. Grossberg, Visual illusions in neural networks: Line neutralization, tilt after-effect, and angle expansion, *J. Theoret. Biol.* 61:477–504 (1976).

138 D. S. Levine and C. D. Woody, Effects of active versus passive dendritic membrane on the transfer properties of a simulated neuron, *Biol. Cybernet.* 31:63–70 (1978).

139 C. von der Malsburg, Self-organization of orientation sensitive cells in the striate cortex, *Kybernetik* 14:85–100 (1973).

140 D. Marr, A theory of cerebellar cortex, *J. Physiol.* 202:437–470 (1969).

141 D. Marr, A theory for cerebral neocortex, *Proc. Roy. Soc. London Ser. B* 176:161–234 (1970).

142 D. Marr, Simple memory: A theory for archicortex, *Philos. Trans. Roy. Soc. London.* 262:23–81 (1971).

143 D. Marr, *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*, Freeman, San Francisco, 1982.

144 D. Marr and T. Poggio, A cooperative computation of stereo disparity, *Science* 194:283–287 (1976).

145 D. Marr and T. Poggio, A computational theory of human stereo vision, *Proc. Roy. Soc. London Ser. B* 204:301–328 (1979).

146 W. S. McCulloch, *Embodiments of Mind*, MIT Press, Cambridge, Mass., 1965.

147 W. S. McCulloch and W. Pitts, A logical calculus of the ideas immanent in nervous activity, *Bull. Math. Biophys.* 5:115–133.

148 N. E. Miller, Some reflections on the law of effect produce a new alternative to drive reduction, in *Nebraska Symposium on Motivation* (M. R. Jones, Ed.), Univ. of Nebraska Press, Lincoln, 1963, pp. 65–112.

149 M. Minsky and S. Papert, *Perceptrons: An Introduction to Computational Geometry*, MIT Press, Cambridge, Mass., 1969.

150 F. S. Montalvo, Consensus versus competition in neural networks, *Internat. J. Man-Mach. Stud.* 7:333–346 (1975).

151 I. Morishita and A. Yajima, Analysis and simulation of networks of mutually inhibiting neurons, *Kybernetik* 11:154–165 (1972).

152 F. Morrell, Effect of anodal polarization on the firing pattern of single cortical cells, *Ann. N.Y. Acad. Sci.* 92:860–876 (1961).
153 G. Moruzzi and H. W. Magoun, Brain stem reticular formation and activation of the EEG, *EEG Clin. Neurophys.* 1:445–473 (1949).
154 V. B. Mountcastle, Modality and topographic properties of single neurons of cat's somatic sensory cortex, *J. Neurophys.* 20:408–434 (1957).
155 V. B. Mountcastle, J. C. Lynch, A. Gergopoulos, H. Sakata, and C. Akuna, Posterior parietal association cortex of the monkey: Command functions for operations within extrapersonal space, *J. Neurophys.* 38:871–908 (1975).
156 V. B. Mountcastle, R. A. Anderson, and B. C. Motter, The influence of attentive fixation upon the excitability of light-sensitive neurons of the posterior parietal cortex, *J. Neurosci.* 1:1218–1235 (1981).
157 J. Nagumo, S. Arimoto, and S. Yoshizawa, An active pulse transmission line simulation nerve axon, *Proc. IRE* 50:2061–2070 (1962).
158 M. M. Nass and L. N. Cooper, A theory for the development of feature detecting cells in the visual cortex, *Biol. Cybernet.* 19:1–18 (1975).
159 J. von Neumann, Probabilistic logics and the synthesis of reliable organisms from unreliable components, in *Cerebral Mechanisms in Behavior: The Hixon Symposium* (L. A. Jeffress, Ed.), Wiley, New York, 1951.
160 J. von Neumann, *The Brain and the Computer*, Yale U.P., New Haven, Conn., 1958.
161 A. Newell, Harpy, production schemes, and human cognition, in *Perception and Production of Fluent Speech* (R. Cole, Ed.), Erlbaum, Hillsdale, N.J., 1980, p. 290.
162 J. O. O'Keefe and L. Nadel, *The Hippocampus as a Cognitive Map*, Clarendon Press, Oxford, England, 1978.
163 J. Olds, Physiological mechanisms of reward, in *Nebraska Symposium on Motivation* (M. Jones, Ed.), Univ. of Nebraska Press, Lincoln, 1955, pp. 73–138.
164 J. Olds, *Drives and Reinforcements: Behavioral Studies of Hypothalamic Functions*, Raven Press, New York, 1977.
165 C. E. Osgood, *Method and Theory in Experimental Psychology*, Oxford U.P., New York, 1953.
166 K. J. Overton and M. A. Arbib, The extended branch-arrow model of the formation of retino-tectal connections, *Biol. Cybernet.* 45:157–175 (1982).
167 J. M. Pearce and G. Hall, A model for Pavlovian learning: Variations in the effectiveness of conditioned but not unconditioned stimulus, *Psych. Rev.* 87:532–552 (1980).
168 A. Pellionisz and R. Llinas, Brain modeling by tensor network theory and computer simulation. The cerebellum: Distributed processor for predictive computation, *Neuroscience* 4:323–348 (1979).
169 A. Pellionisz and R. Llinas, Tensorial approach to the geometry of brain function: Cerebellar coordination via a metric tensor, *Neuroscience* 5:1125–1136 (1980).
170 A. Pellionisz and R. Llinas, Space-time representation in the brain. The cerebellum as a predictive space-time metric tensor, submitted to *Neuroscience.*
171 R. Perez, L. Glass, and R. Shlaer, Development of specificity in the cat visual cortex, *J. Math. Biol.* 1:275–288 (1975).
172 D. H. Perkel and T. H. Bullock, Neural coding, *Neurosci. Res. Prog. Bull.* 6:221–347 (1968).
173 J. D. Pettigrew and T. Kasamatsu, Local perfusion of noradrenaline maintains visual cortical plasticity, *Nature* 271:761–763 (1978).

174 W. T. Powers, *Behavior: The Control of Perception*, Aldine, Chicago, 1973.
175 I. Prigogine, Structure, dissipation, and life, in *First International Conference, Theoretical Physics and Biology, Versailles*, North-Holland, Amsterdam, 1969, pp. 23–52.
176 W. Rall, Electrophysiology of a dendritic neuron model, *Biophys. J.* 2:145–167 (1962).
177 W. Rall, Distinguishing theoretical synaptic potentials computed for different soma-dendritic distributions of synaptic input, *J. Neurophys.* 30:1138–1168 (1967).
178 A. Rapoport, Ignition phenomena in random nets, *Bull. Math. Biophys.* 14:35–44 (1952).
179 N. Rashevsky, *Mathematical Biophysics*, Dover, New York, 1960, Vol. II.
180 F. Ratliff, *Mach Bands: Quantitative Studies of Neural Networks in the Retina*, Holden-Day, San Francisco, 1965.
181 W. Reichardt and G. MacGinitie, Zur Theorie der lateralen Inhibition, *Kybernetik* 1:155–165 (1962).
182 R. A. Rescorla and A. R. Wagner, A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and non-reinforcement, in *Classical Conditioning II: Current Research and Theory* (A. H. Black and W. F. Prokasy, Eds.), Appleton-Century-Crofts, New York, 1972.
183 D. A. Robinson, Adaptive gain control of vestibulo-ocular reflex by the cerebellum, *J. Neurophys.* 39:954–969 (1976).
184 F. Rosenblatt, *Principles of Neurodynamics*, Spartan Books, Washington, 1962.
185 V. S. Rusinov, An electrophysiological analysis of the connecting function in the cerebral cortex in the presence of a dominant area, in *Abstract of the Nineteenth International Physiological Conference*, Montreal, 1953, pp. 719–720.
186 M. E. Scheibel and A. B. Scheibel, Anatomical basis of attention mechanisms in vertebrate brains, in *The Neurosciences, a Study Program* (G. C. Quarton, T. Melnechuk, and F. O. Schmitt, Eds.), Rockefeller U.P., New York, 1967, pp. 577–602.
187 A. C. Scott, *Neurophysics*, Wiley-Interscience, New York, 1977.
188 P. Scott, Conditional control transfer mechanisms in the neocortex: 1. The basic model, *J. Theoret. Biol.* 81:423–452 (1980).
189 P. Scott, Conditional control transfer mechanisms in the neocortex: 2. Cytoarchitectural and electrophysiological evidence and predictions, *J. Theoret. Biol.* 81:453–472 (1980).
190 T. J. Sejnowski, On the stochastic dynamics of neuronal interaction, *Biol. Cybernet.* 23:203–211 (1976).
191 C. Shannon, A mathematical theory of communication, *Bell Syst. Tech. J.* 27:379–423 (1948).
192 C. Shannon and W. Weaver, *The Mathematical Theory of Communication*, Univ. of Illinois Press, Urbana, 1949; rev. ed., 1964.
193 H. A. Simon, *The Sciences of the Artificial*, MIT Press, Cambridge, Mass., 1969.
194 R. L. Solomon, L. J. Kamin, and L. C. Wynne, Traumatic avoidance learning: The outcomes of several extinction procedures with dogs, *J. Abnorm. Soc. Psych.* 48:291–302 (1953).
195 G. Sperling and M. M. Sondhi, Model for visual luminance discrimination and flicker detection, *J. Opt. Soc. Amer.* 58:1133–1145 (1968).
196 K. C. Squires, C. Wickens, N. K. Squires, and E. Donchin, The effect of stimulus sequence on the waveform of the cortical event-related potential, *Science*

193:1142–1146 (1976).
197 G. S. Stent, A physiological mechanism for Hebb's postulate of learning, *Proc. Nat. Acad. Sci. U.S.A.* 70:997– 103 (1973).
198 S. Sternberg, Memory-scanning: Mental processes revealed by reaction-time experiments, *Amer. Sci.* 57:421–457 (1969).
199 R. S. Sutton and A. G. Barto, Toward a modern theory of adaptive networks: Expectation and prediction, *Psych. Rev.* 88:135–170 (1981).
200 J. Theios, P. G. Smith, S. E. Haviland, J. Traupmann, and M. C. Moy, Memory scanning as a serial self-terminating process, *J. Exp. Psych.* 97:323–336 (1973).
201 R. Thom, *Structural Stability and Morphogenesis*, Benjamin, Reading, Mass., 1975.
202 R. F. Thompson, The search for the engram II, in *Neural Mechanisms and Behavior* (D. McFadden, Ed.), Springer, New York, 1979, pp. 172–222.
203 A. M. Uttley, The transmission of information and the effect of local feedback in theoretical and neural networks, *Brain Res.* 2:21–50 (1966).
204 A. M. Uttley ,in *Proceedings of IEE Conference No. 42*, 1968, pp. 338–346.
205 A. M. Uttley, The informon: A network for adaptive pattern recognition, *J. Theoret. Biol.* 27:31–67 (1970).
206 A. M. Uttley, The informon in classical conditioning, *J. Theoret. Biol.* 49:355–376 (1975).
207 A. M. Uttley, A two-pathway informon theory of conditioning and adaptive pattern recognition, *Brain Res.* 102:23–35 (1976).
208 A. M. Uttley, Simulation studies of learning in an informon network, *Brain Res.* 102:35–53 (1976).
209 A. M. Uttley, Neurophysiological predictions of a two-pathway informon theory of neural conditioning, *Brain Res.* 102:55–70 (1976).
210 N. Wiener, *Cybernetics*, Wiley, New York, 1948.
211 D. J. Willshaw and C. von der Malsburg, How patterned neural connections can be set up by self-organization, *Proc. Roy. Soc. London Ser. B* 194:431–445 (1976).
212 H. R. Wilson, A synaptic model for spatial frequency adaptation, *J. Theoret. Biol.* 50:327–352 (1975).
213 H. R. Wilson, A transducer function for threshold and suprathreshold human vision, *Biol. Cybernet.* 38:171–178 (1980).
214 H. R. Wilson and J. D. Cowan, Excitatory and inhibitory interactions in localized populations of model neurons, *Biophys. J.* 12:1–24 (1972).
215 H. R. Wilson and J. D. Cowan, A mathematical theory of the functional dynamics of cortical and thalamic nervous tissue, *Kybernetik* 13:55–80 (1973).
216 R. Wong and E. Harth, Stationary states and transients in neural populations *J. Theoret. Biol.* 40:77–106 (1973).
217 C. D. Woody, *Memory, Learning, and Higher Function: A Cellular View*, Springer, New York, 1982.
218 C. D. Woody, A. A. Buerger, R. A. Ungar, and D. S. Levine, Modeling aspects of learning by altering biophysical properties of a simulated neuron, *Biol. Cybernet.* 23:73–82 (1976).
219 C. C. Zeeman, *Catastrophe Theory: Selected Papers*, Addison-Wesley, Reading, Mass., 1977.

The Flux-Ratio Equation under Nonstationary Boundary Conditions

L. BASS AND A. J. BRACKEN
Department of Mathematics, University of Queensland, St. Lucia, 4067, Brisbane, Australia

Received 4 October 1982; revised 17 February 1983

ABSTRACT

Ussing's equation for the ratio of unidirectional tracer fluxes, passing in opposite directions through a membrane which has transport properties varying with the distance from a boundary, has recently been extended to include nonstationary fluxes under stationary boundary conditions [Sten-Knudsen and Ussing, *J. Membrane Biol.* 63:233–242 (1981)]. Using a concise and more generally valid method of proof, that extension is generalized further to include nonstationary boundary values of the tracer concentrations, and to obtain new results useful in the design of experiments.

Ussing's classic result concerning the ratio of stationary unidirectional tracer fluxes has been extended profoundly by Sten-Knudsen and Ussing [1] to include nonstationary fluxes under stationary boundary values of the concentrations. This result can be extended still further to include nonstationary boundary conditions, with the help of a simpler and more generally valid method of proof, as we now show.

A membrane is represented by a slab occupying the spatial interval $0 \leqslant x \leqslant h$, in which a tracer has diffusion coefficient $D(x)$ and drift velocity $v(x)$. Both D and v may vary arbitrarily with x, except that $D(x) > 0$ for all x, but they are both assumed independent of the time t. The concentration $c(x, t)$ and flux $j(x, t)$ of the tracer in the slab satisfy the continuity equation

$$\frac{\partial c}{\partial t} + \frac{\partial j}{\partial x} = 0 \tag{1}$$

and the flux equation

$$j = -D(x)\frac{\partial c}{\partial x} + v(x)c. \tag{2}$$

©Elsevier Science Publishing Co., Inc., 1983
52 Vanderbilt Ave., New York, NY 10017 0025-5564/83/$03.00

There is no tracer in the slab until $t = 0$, when some given positive concentration is imposed at one of the two boundaries, while $c = 0$ is maintained at the other boundary. We denote the two resulting cases by the subscripts 1,2: When $c_1(0,t) > 0$ and $c_1(h,t) = 0$, the concentration is $c_1(x,t)$, the flux is $j_1(x,t)$, and the unidirectional outflux from the slab is $j_1(h,t) > 0$. When $c_2(h,t) > 0$ and $c_2(0,t) = 0$, the unidirectional outflux from the slab is $j_2(0,t) < 0$. Sten-Knudsen and Ussing [1] showed that if $c_1(0,t) = c_1(0)$ and $c_2(h,t) = c_2(h)$ are independent of time, then at all times $t > 0$

$$\frac{j_1(h,t)}{-j_2(0,t)} = \frac{c_1(0)}{c_2(h)} \exp\left(\int_0^h \frac{v}{D}\,dx\right); \tag{3}$$

this flux ratio is constant and therefore equal to the ratio $-j_1(h,\infty)/j_2(0,\infty)$ in the ultimate steady state, when Equation (3) is reduced to Ussing's classic relation. (The negative sign of j_2 is commonly omitted.)

In what follows we generalize this result further by proving that Equation (3) remains valid if $c_1(0)$ and $c_2(h)$ are replaced by boundary values $c_1(0,t)$ and $c_2(h,t)$ which may vary arbitrarily with time, provided that their ratio r does not depend on time:

$$r = \frac{c_1(0,t)}{c_2(h,t)} = \text{const.} \tag{4}$$

Practical studies of flux ratios may benefit from this further relaxation of the requisite experimental conditions: in place of the strict constancy of $c_1(0)$ and $c_2(h)$ at all $t > 0$, emphasized by Sten-Knudsen and Ussing [1], the boundary concentrations now admit one arbitrary function of time which the experimentalist may choose at will.

In proving this result [which includes Equation (3)] we shall not follow Sten-Knudsen and Ussing [1] in representing $D(x)$ and $v(x)$ as piecewise constant functions each with a finite number of jumps (that is, we shall not compose the slab from a finite number of homogeneous laminae). Such a representation, while permitting the construction of explicit solutions, would leave unanswered questions concerning the validity of the proof when $D(x)$ and $v(x)$ vary more generally. In particular, the term $c\,dv/dx$, obtained when Equation (2) is inserted in Equation (1), would be zero except possibly at a finite number of discrete x-values. [It is the term $c\,dv/dx$ that prevents $c = \text{const} \neq 0$ from being a solution of Equations (1) and (2). For a charged tracer, dv/dx depends linearly on the local charge density associated with the membrane.]

One may wonder what can be deduced in the still more general case when the boundary values $c_1(0,t)$, $c_2(h,t)$ are not related as in Equation (4): we discuss this briefly in the Appendix.

We introduce the new dependent variable

$$u(x,t) = c(x,t)\exp\left(-\int_0^x \frac{v}{D}dx\right). \tag{5}$$

We note that this transformation would be inadmissible if $D(x)$ vanished in some interval $x, x+\Delta x$ in which v is nonzero (positive or negative): the tracer would drift across that interval in one direction but not in the other, so that for all $t>0$ either $j_1(h,t)$ or $j_2(0,t)$ would be zero, and the flux ratio would be permanently zero or infinity. In what follows we assume that $D>0$, and that the integral of v/D exists across any part of the slab.

In terms of $u(x,t)$, Equation (1) becomes

$$\frac{\partial u}{\partial t}\exp\left(\int_0^x \frac{v}{D}dx\right) + \frac{\partial j}{\partial x} = 0 \tag{6}$$

and Equation (2) becomes

$$j = -D\frac{\partial u}{\partial x}\exp\left(\int_0^x \frac{v}{D}dx\right). \tag{7}$$

The initial condition is $c(x,0)=0$ or, from Equation (5),

$$u(x,0) = 0. \tag{8}$$

We shall use Laplace transforms of u and j; thus,

$$\bar{u}(x,p) = \int_0^\infty e^{-pt}u(x,t)\,dt, \tag{9}$$

and similarly for the transform $\bar{j}$ of j. Using also Equation (8), the transforms of Equations (6) and (7) are

$$p\bar{u}\exp\left(\int_0^x \frac{v}{D}dx\right) + \frac{d\bar{j}}{dx} = 0, \tag{10}$$

$$\bar{j} = -D\frac{d\bar{u}}{dx}\exp\left(\int_0^x \frac{v}{D}dx\right). \tag{11}$$

We now consider any two solutions of Equations (6), (7), and (8), say (u_1, j_1) and (u_2, j_2), which satisfy different sets of boundary conditions. Their Laplace transforms are $(\bar{u}_1, \bar{j}_1)$ and $(\bar{u}_2, \bar{j}_2)$. Envisage writing Equation (10) first for $(\bar{u}_1, \bar{j}_1)$, and then again for $(\bar{u}_2, \bar{j}_2)$. We multiply the first version of Equation (10) through with $\bar{u}_2$, the second version with $\bar{u}_1$, and subtract.

The terms with $p\bar{u}_1\bar{u}_2$ cancel out, and we obtain

$$0 = \bar{u}_2\frac{d\bar{j}_1}{dx} - \bar{u}_1\frac{d\bar{j}_2}{dx} = \frac{d}{dx}(\bar{u}_2\bar{j}_1 - \bar{u}_1\bar{j}_2) - \left(\bar{j}_1\frac{d\bar{u}_2}{dx} - \bar{j}_2\frac{d\bar{u}_1}{dx}\right), \tag{12}$$

where the second equality is an identity. Using $\bar{j}_1$ and $\bar{j}_2$, respectively, as given by Equation (11), the last bracket on the right-hand side of Equation (12) vanishes, enabling us to conclude that $\bar{u}_2\bar{j}_1 - \bar{u}_1\bar{j}_2$ is independent of x. In particular, at the boundaries,

$$\bar{u}_2(0)\bar{j}_1(0) - \bar{u}_1(0)\bar{j}_2(0) = \bar{u}_2(h)\bar{j}_1(h) - \bar{u}_1(h)\bar{j}_2(h). \tag{13}$$

We now make use of the two sets of boundary conditions for $t > 0$, described at the outset, which determine the two opposite unidirectional fluxes. Firstly,

$$c_1(0,t) = f(t) > 0, \qquad c_1(h,t) = 0, \tag{14}$$

where $f(t)$ is arbitrary except for being positive and finite. Using Equation (5), Equation (14) becomes

$$u_1(0,t) = f(t), \qquad u_1(h,t) = 0, \tag{15}$$

with the transforms

$$\bar{u}_1(0) = \bar{f}, \qquad \bar{u}_1(h) = 0. \tag{16}$$

Secondly, using also Equation (4),

$$c_2(0,t) = 0, \qquad c_2(h,t) = \frac{c_1(0,t)}{r} = \frac{f(t)}{r}, \tag{17}$$

so that, using Equation (5),

$$u_2(0,t) = 0, \qquad u_2(h,t) = \frac{f(t)}{r}\exp\left(-\int_0^h \frac{v}{D}dx\right), \tag{18}$$

and on transforming,

$$\bar{u}_2(0) = 0, \qquad \bar{u}_2(h) = \frac{\bar{f}}{r}\exp\left(-\int_0^h \frac{v}{D}dx\right). \tag{19}$$

Now, using Equations (16) and (19) in Equation (13), we obtain

$$-\bar{f}\bar{j}_2(0) = \frac{\bar{f}\bar{j}_1(h)}{r}\exp\left(-\int_0^h \frac{v}{D}dx\right). \tag{20}$$

Since $f(t)$ is finite and positive, $\bar{f}(p)$ is finite and positive for all $p > 0$. We can therefore cancel $\bar{f}$ in Equation (20) and obtain

$$\bar{j}_1(h) = -r\bar{j}_2(0)\exp\left(\int_0^h \frac{v}{D}dx\right). \tag{21}$$

Since Equation (21) for the Laplace transforms $\bar{j}_1(h)$ and $\bar{j}_2(0)$ holds for all $p > 0$, it holds for $j_1(h,t)$ and $j_2(0,t)$ themselves by the uniqueness (Lerch) theorem for Laplace transforms. Hence

$$\frac{j_1(h,t)}{-j_2(0,t)} = r\exp\left(\int_0^h \frac{v}{D}dx\right), \tag{22}$$

which, in view of Equation (4), is the desired result.

APPENDIX

Consider in place of Equations (14) and (17) the boundary conditions

$$\begin{aligned} c_1(0,t) &= f_1(t) > 0, & c_1(h,t) &= 0 \\ c_2(0,t) &= 0, & c_2(h,t) &= f_2(t) > 0, \end{aligned} \tag{23}$$

with f_1, f_2 possibly different functions, so that in general Equation (4) no longer holds. Then we obtain in place of Equation (20)

$$-\bar{f}_1\bar{j}_2(0) = \bar{f}_2\bar{j}_1(h)\exp\left(-\int_0^h \frac{v}{D}dx\right) \tag{24}$$

with $\bar{f}_1, \bar{f}_2$ the transforms of f_1, f_2. Since the product of two Laplace transforms is the transform of a convolution, Equation (24) implies that, at all times $t > 0$,

$$\frac{\int_0^t c_2(h,t-\tau)\,j_1(h,\tau)\,d\tau}{-\int_0^t c_1(0,t-\tau)\,j_2(0,\tau)\,d\tau} = \exp\left(\int_0^h \frac{v}{D}dx\right). \tag{25}$$

This result is the generalization of Equation (22) appropriate to the boundary conditions (23).

Special results which may be useful in the design of experiments can be obtained by choosing f_1 and f_2 so that their Laplace transforms are simply related. For example, consider the case when one of the boundary values is constant, and the other rises from zero linearly with the time:

$$\begin{aligned} c_1(0,t) &= \alpha t, & c_1(h,t) &= 0 \\ c_2(0,t) &= 0, & c_2(h,t) &= c_2(h), \end{aligned} \tag{26}$$

where α and $c_2(h)$ are positive constants. Then Equation (24) leads, in place of Equation (25), to

$$\frac{j_1(h,t)}{-\int_0^t j_2(0,\tau)\,d\tau} = \frac{\alpha}{c_2(h)} \exp\left(\int_0^h \frac{v}{D}\,dx\right). \tag{27}$$

Thus in this case the ratio of the instantaneous outflux at one boundary to the accumulated outflux at the other is independent of time.

One of us (L.B.) is obliged to Professor O. Sten-Knudsen for a discussion which stimulated the present work; and to Professor C. Crone and to his Institute for the hospitality which led to that discussion.

REFERENCES

1. O. Sten-Knudsen and H. H. Ussing, The flux ratio equation under nonstationary conditions, *J. Membrane Biol.* 63:233–242 (1981).

The Concentration of Modified Protein Species as an Index of Protein Modification Cooperativity

EMMANUEL T. RAKITZIS
Department of Biological Chemistry, University of Athens Medical School, Athens 620, Greece

Received 20 November 1982; revised 12 April 1983

ABSTRACT

A mathematical treatment of stoichiometric protein modification reactions is presented, and a formula is derived for the determination of the concentration of the *i*-fold modified protein species as a function of the concentration of the unreacted protein moiety. The method is applied to a case from the literature: the carbamylation of amino-terminal valine residues of human hemoglobin.

Protein modification cooperativity (irreversible inhibitor binding cooperativity) is the sequential modification of a protein in such a manner that the partially modified protein species possess different reactivity towards the modifying agent than does the unmodified moiety. Stated differently, protein modification cooperativity is the sequential modification of a protein in a manner different than can be predicted by the theory of probability alone [1–4]. In protein modification cooperativity the concentration of protein modified residues is described by a summation of exponential functions of the reaction time, graphical analysis of which may yield the values for the different microscopic modification rate constants [2–4]. In this communication it is shown that the concentration of the different modified protein species may also be used to probe into the nature of the sequential modification of a protein, viz., to answer the question whether the sequential modification of a protein is a process involving cooperative dependence of the separate modification steps. It is seen that the method has a wider applicability than the time dependence of the modification of the protein reactive residues. This is so because protein modification is described by linear differential equations, with modified protein residue concentration as the dependent variable and reaction time as the independent variable, only when the concentration of the modifying agent is adequately high; at lower concentrations of the modifying agent the differential equations describing

©Elsevier Science Publishing Co., Inc., 1983
52 Vanderbilt Ave., New York, NY 10017 0025-5564/83/$03.00

this process are nonlinear. The method described in this communication is also applicable when the differential equation describing the rate of change in the concentration of each protein species, with time as the independent variable, is a second order equation.

A protein with n intrinsically identical modifiable residues per protein molecule is considered. The stepwise stoichiometric modification of this protein, to produce fully modified protein, may be written as follows:

$$E_n \overset{k_n}{\rightarrow} E_{n-1} \overset{k_{n-1}}{\rightarrow} E_i \overset{k_i}{\rightarrow} E_1 \overset{k_1}{\rightarrow} E_0, \tag{1}$$

where E_i is the protein species possessing i moles of unmodified residues per mole of protein, and k_i is the rate constant for the modification of this species, to produce the species with $i-1$ moles of unmodified residues per mole of protein. In the absence of protein modification cooperativity $k_i = ik_1$:

The rate of change of $[E]_i$ is

$$\frac{d[E]_i}{dt} = k_{i+1}[E]_{i+1} - k_i[E]_i. \tag{2}$$

The rate of change of $[E]_n$ is

$$\frac{d[E]_n}{dt} = -k_n[E]_n. \tag{3}$$

Writing out Equation (2) for $i = n-1$, and dividing this equation by Equation (3), we have

$$\frac{d[E]_{n-1}}{d[E]_n} = \frac{k_{n-1}}{k_n}\frac{[E]_{n-1}}{[E]_n} - 1. \tag{4}$$

Equation (4), a Bernoulli equation, may be solved for $[E]_{n-1}$, and the process may be repeated to give the fractional concentration of the species E_i:

$$\frac{[E]_i}{[E]_{n(0)}} = \sum_{j=0}^{n-1} C_{(i)j}[E]_n^{k_{n-j}/k_n}$$

$$= \frac{k_{i+1}}{k_n}\sum_{j=0}^{n-1} C_{(i+1)j}\frac{\left(\frac{[E]_n}{[E]_{n(0)}}\right)^{k_i/k_n} - \left(\frac{[E]_n}{[E]_{n(0)}}\right)^{k_{n-j}/k_n}}{\frac{k_{n-j}-k_i}{k_n}[E]_{n(0)}^{-(k_{n-j}-k_n)/k_n}}, \tag{5}$$

where $[E]_{n(0)}$ is $[E]_n$ at $t = 0$, and also where $C_{(n)0} = k_n/(k_{n-1} - k_n)$. It will

be noted that $k_1,\ldots,k_n$ are only apparent first order constants, since these "constants" are functions of the modifying agent M. It will be seen, however, that if the dependence of $k_1,\ldots,k_n$ on $[M]$ is such that $k_i = k_{i(0)}[M]$ (where $k_{i(0)}$ is the second order modification rate constant), then the values k_{n-1}/k_n, and k_{n-j}/k_n, in Equations (4) and (5) reduce to $k_{n-1(0)}/k_{n(0)}$, and $k_{n-j(0)}/k_{n(0)}$, i.e., they reduce to values that are free of $[M]$. Accordingly, Equation (5) is also applicable at low modifying agent concentrations, at which Equations (2) and (3) describe second order reactions.

The treatment of stoichiometric protein modification cooperativity outlined in this communication may be applied to the carbamylation of the amino-terminal valine residues of human hemoglobin [5]. CO-hemoglobin was carbamylated by reaction with 0.1 M [^{14}C]KNCO in a 0.01 M sodium phosphate buffer, pH 6.5, and the stoichiometric carbamylated hemoglobin species were separated into lysine-reacted and lysine-unreacted species, by means of DEAE- and CM-cellulose column chromatography [5]. Since the number of moles of carbamylated valine per mole of hemoglobin was the same in the lysine-reacted and lysine-unreacted species, it is permissible to consider the former species to have the same composition in stoichiometric valine-modified hemoglobin species as the latter (lysine-unreacted) species.

The valine-modified hemoglobin species could be distinguished into the unreacted hemoglobin moiety, the dimodified species, and the tetramodified species. The dimodified and the tetramodified hemoglobin species were also separated into alpha and beta subunits. As would be expected in stoichiometric protein modification, the alpha and beta subunits were symmetrically labeled. The fractional concentration of the unmodified, dimodified, and tetramodified hemoglobin species was calculated from Figure 3 of [5] by cutting out the areas under each peak, and weighing the pieces of paper so produced. The values obtained are given in Table 1 against the calculated values for each hemoglobin species, in the absence of modification cooperativity. Calculation of the value of each stoichiometric modifed hemoglobin species, in the absence of modification cooperativity, was by the use of Equation (5) and the substitution of ik_1 for k_i. It will be seen from Table 1 that the value of the experimentally determined tetramodified hemoglobin species is one order of magnitude larger than the calculated value. This indicates the presence of positive protein modification cooperativity [2]. The value for the dimodified species is smaller than the calculated value, while the monomodified as well as the trimodified species could not be identified in the column eluate [5]. The lack of determinable values for each stoichiometric modified hemoglobin species precludes a quantitative application of the method presented here, viz., the determination of each stoichiometric modification rate constant. It should be possible to successfully apply the method of diagnosis of stoichiometric protein modification cooperativity outlined in this communication, to data obtained by means of natural abundance ^{13}C nuclear

TABLE 1

Calculated and Experimentally Found Values of Stoichiometric Carbamylated Hemoglobin Species

Carbamylated Hb species[a]	Calculated values of Hb species in the absence of cooperative modification: As a function of the fractional concentration of unmodified Hb	For a fractional concentration of unmodified Hb of 0.0822	Experimentally found values of Hb species[b]
$[Hb]_4/[Hb]_{4(0)}$	a	0.0822	0.0822
$[Hb]_3/[Hb]_{4(0)}$	$-4a+4a^{3/4}$	0.2850	0
$[Hb]_2/[Hb]_{4(0)}$	$6a-12a^{3/4}+6a^{2/4}$	0.3710	0.298
$[Hb]_1/[Hb]_{4(0)}$	$-4a+12a^{3/4}-12a^{2/4}+4a^{1/4}$	0.2130	0
$[Hb]_0/[Hb]_{4(0)}$	$a-4a^{3/4}+6a^{2/4}-4a^{1/4}+1$	0.0480	0.619

[a]For a definition of the notation used, see Equation (1). $[Hb]_{4(0)}$ is the concentration of unmodified hemoglobin at the start of the reaction. Accordingly, the values given correspond to the fractional concentrations of unmodified, monomodified, dimodified, trimodified, and tetramodified hemoglobin, in the order given.

[b]Values calculated from the data of Williams, Chung, and Schuster [5]. For details of the method used in the calculation, see the text.

magnetic resonance spectroscopy, since by this method a determination of the fractional concentration of protein modified species is possible in the unfractionated homogenate preparation [6–7].

REFERENCES

1 W. J. Ray and D. E. Koshland, Jr., A method characterizing the type and numbers of groups involved in enzyme action, *J. Biol. Chem.* 236:1973–1979 (1961).

2 E. T. Rakitzis, Kinetics of irreversible enzyme inhibition: Co-operative effects, *J. Theoret. Biol.* 67:49–59 (1977).

3 E. T. Rakitzis, Kinetics of irreversible enzyme inhibition by an unstable inhibitor: Co-operative effects, *J. Theoret. Biol.* 75:239–243 (1978).

4 E. T. Rakitzis, Kinetic analysis of biphasic protein modification reactions, *J. Math. Biol.* 10:79–87 (1980).

5 R. C. Williams, Jr., L. L. Chung, and T. M. Schuster, Preparation and purification of carbamylated intermediates of human hemoglobin, *Biochem. Biophys. Res. Commun.* 62:118–128 (1975).

6 W. J. Goux and A. Allerhand, Studies of chemically modified histidine residues of proteins by carbon 13 nuclear magnetic resonance spectroscopy, *J. Biol. Chem.* 254:2210–2213 (1979).

7 R. S. Norton and A. Allerhand, Studies of chemical modifications of proteins by carbon 13 nuclear magnetic resonance spectroscopy, *J. Biol. Chem.* 251:6522–6528 (1976).

The Role of Complexity in Comparing Classifications*

WILLIAM H. E. DAY
Department of Computer Science, Memorial University of Newfoundland, St. John's, Newfoundland, Canada AlC 5S7

Received 19 December 1982; revised 15 April 1983

ABSTRACT

In numerical taxonomy there is considerable interest in developing theory and methodology to provide for the quantitative comparison of classifications. I describe a mathematical model in which this problem can be addressed, and with it I develop a quantitative measure of the fit of a consensus classification to the set of classifications from which the consensus is derived. The model assumes that classifications are related by a partial order and have associated with them a quantitative measure of classification complexity. I use basic results concerning valuations on partially ordered sets to exhibit relationships among the concepts of classification complexity, distance between classifications, and strict consensus of classifications. Classification complexity is extended parsimoniously to provide a measure of the complexity of *sets* of classifications. Computing this set complexity measure is a difficult, in fact NP-complete, problem. The measure of consensus fit is based on Papentin's idea that pattern complexity can be differentiated into organized complexity (i.e., the minimal description of rules underlying the pattern) and unorganized complexity (i.e., the minimal description of the random aspects of the pattern). Imagine a set of classifications as forming a pattern within the poset of classifications: the complexity of a consensus classification estimates the pattern's organized complexity, while its unorganized complexity is estimated by a variation of the set complexity problem. These estimates are used to construct a normalized fit measure whose extreme values are approached as the corresponding estimates approach zero. The fit measure is illustrated using majority-rule and strict-consensus methods. Computing this fit measure is a difficult, in fact NP-complete, problem.

1. INTRODUCTION

In recent years there has been considerable interest in developing theory and methodology for the evaluation of taxonomic methods, much of it stimulated, no doubt, by the desire to evaluate the extent to which taxonomic methods satisfy the goals of various schools of classification (see Mickevich and Farris [23], Rohlf and Sokal [33], and references therein). Being addressed are important issues involving taxonomic congruence (addition of

*This work was supported in part by the Natural Sciences and Engineering Research Council of Canada under Grant A-4142.

©Elsevier Science Publishing Co., Inc., 1983
52 Vanderbilt Ave., New York, NY 10017 0025-5564/83/$03.00

characters to a classification) [13, 21, 22, 32, 33, 35, 36, 39], stability (addition of operational taxonomic units to a classification) [33, 36, 39], character state predictivity [4, 33], and naturalness [11–13, 33] of classifications. Inherent in the resolution of such issues is a need to obtain quantitative measures of the agreement of classifications.

The problem, of course, is not new, and it has been approached from at least two points of view. In the first, one compares two classifications using a quantitative measure of their dissimilarity (or similarity) [3, 6, 7, 8, 10, 30, 31, 38]; frequently such dissimilarity measures exhibit the metric properties and so may be interpreted as measures of distance. In the second, one derives from a given set of classifications a consensus classification that in some sense best represents the members of the given set [1, 5, 18–20, 27, 31, 37].

These two approaches are not, nor should they be, entirely unrelated. A consensus can be defined to minimize the value of an appropriate function of the distances between it and the classifications in the given set. When the set contains exactly two classifications, a consensus may in fact lie on a path of minimum length connecting the two classifications. Relationships between distance and consensus can also be understood in terms of a measure of classification structure or complexity. In taxonomy this concept has intuitive appeal: usually one can identify the least complex, or trivial, classification; and when comparing two classifications there may be plausible reasons to identify one as being more, or less, complex than the other. But the significance of classification complexity lies beyond its intuitive appeal: distance between two classifications can be a quantitative measure of the difference in their complexities; and the complexity of a consensus can be a quantitative measure of the structure exhibited (in some sense) by the classifications in the original set.

Thus it seems desirable that taxonomists understand not only the concepts of distance, consensus, and complexity, but the relationships that exist among them. To this end I formulate a mathematical model in which such relationships can be explored, and I illustrate the model's utility by developing a measure of the fit of a consensus to the set of classifications on which it is based. In Section 2 I define basic concepts and formulate the model as a partially ordered set of classifications. In Section 3 I use basic results in poset theory to establish relationships among the concepts of distance, complexity, and strict consensus. Within the model, complexity and distance play equivalent roles, since each can be derived from the other. I make explicit the assumptions underlying the model, so that taxonomists can assess its relevance to a given application. In Section 4 I generalize the concept of classification complexity to apply as well to *sets* of classifications. Set complexity is based on a most parsimonious description of the set within the poset of classifications. Computing set complexity is a difficult, in fact NP-complete, problem.

Finally, in Section 5 I formulate a quantitative measure of the fit of a consensus to the set of classifications on which it is based. The measure is based on Papentin's idea [28, 29] that pattern complexity can be differentiated into organized complexity (i.e., the minimal description of the rules underlying the pattern) and unorganized complexity (i.e., the minimal description of the random aspects of the pattern). Imagine a set of classifications as forming a pattern within the poset of classifications: the complexity of a consensus estimates the pattern's organized complexity, while its unorganized complexity is estimated by a variation of the set complexity problem. These estimates are used to construct a normalized fit measure whose extreme values are approached as the corresponding estimates approach zero. The fit measure is illustrated by examples using the majority-rule and strict-consensus methods. Computing this fit measure is a difficult, in fact NP-complete, problem.

2. BASIC CONCEPTS

This section introduces the basic elements of a model for comparison of classifications. Complexity, distance, and consensus are defined as functions on a partially ordered set of classifications.

Let X be a set of classifications of interest. Such classifications may be hierarchic or nonhierarchic. Consider, for example, the set P_n of all partitions of the set $N = \{1, \ldots, n\}$ of operational taxonomic units (OTUs). The partition $\{\{1,2,3\},\{4,5\}\}$ in P_5 represents a nonhierarchic classification that contains two clusters, $\{1,2,3\}$ and $\{4,5\}$, of OTUs. Consider, for another example, the set T_n of all n-trees on N. An *n-tree on N* is a set x of subsets of N satisfying: $\varnothing \notin x$; $N \in x$; $\{i\} \in x$ for every $i \in N$; and $A \cap B \in \{\varnothing, A, B\}$ for all $A, B \in x$ [18]. As illustrated in Figure 1, each n-tree in T_n represents a hierarchic classification whose clusters are precisely the elements of the n-tree. P_n and T_n exhibit properties of finiteness and nontriviality that simplify the following discussion, and so I adopt them as properties of X.

(A1) X is a finite nonempty set.

In Figure 1, u_1 is of particular interest, since its hierarchical classification is characterized by the absence of hierarchical structure: u_1 is said to be "trivial" or "completely unresolved." On the other hand, u_3 is "fully resolved," since the hierarchical structure represented by clusters 12, 123, and 45 is (in one sense) the maximum possible in 5-trees. Finally, u_2 lies between these structural extremes: it is more complex than u_1, since it is obtained from u_1 by the insertion of 123; it is less complex than u_3, since it is obtained from u_3 by the deletion of 12 and 45. This ordering of 5-trees by structural relationship is made precise in the following way. Let $\leqslant$ denote the binary relation on T_5 such that for all $x, y \in T_5$, $x \leqslant y$ if and only if $x \subseteq y$; to use the

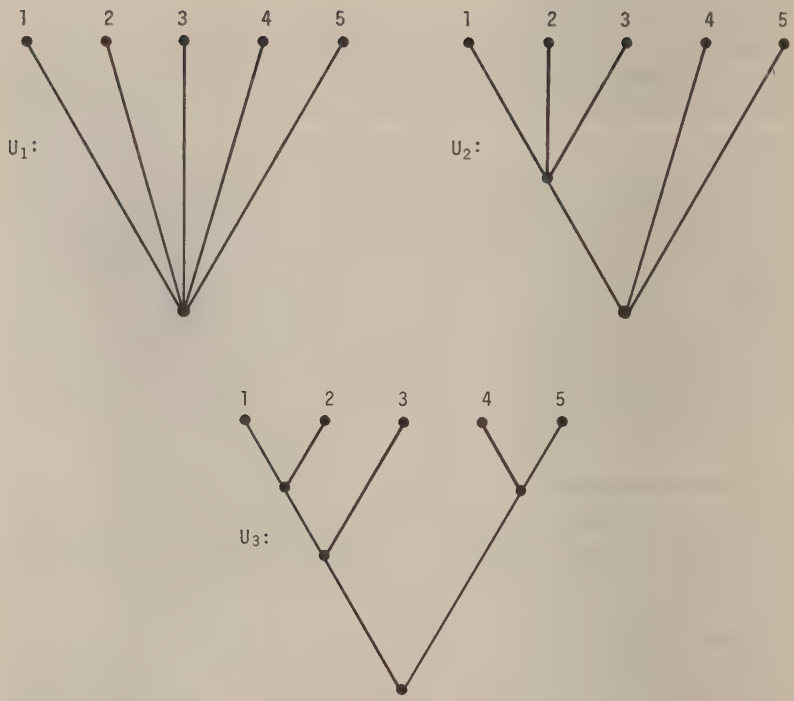

FIG. 1. Examples of 5-trees. If each subset $\{i_1, i_2, \ldots, i_k\}$ of $\{1, \ldots, 5\}$ is denoted by $i_1 i_2 \cdots i_k$, then $u_1 = \{1, 2, 3, 4, 5, 12345\}$, $u_2 = \{1, 2, 3, 4, 5, 123, 12345\}$, and $u_3 = \{1, 2, 3, 4, 5, 12, 45, 123, 12345\}$.

terminology of Estabrook and McMorris [40], $x \leqslant y$ if and only if y is a *refinement* of x. The reader can verify that $\leqslant$ is *reflexive* (i.e., $x \leqslant x$ for all $x \in T_5$), *antisymmetric* (i.e., $x \leqslant y$ and $y \leqslant x$ implies $x = y$ for all $x, y \in T_5$), and *transitive* (i.e., $x \leqslant y$ and $y \leqslant z$ implies $x \leqslant z$ for all $x, y, z \in T_5$), and so $\leqslant$ is a *partial order* on T_5. Observe that u_1 occupies a special position in the poset $(T_5, \leqslant)$, since $u_1 \leqslant x$ for every $x \in T_5$; it is the unique least element, or *zero*, of T_5 and is usually denoted by the symbol 0.

Analogous comments apply to P_5 as well. The partition $\{12345\}$ is completely unresolved, since the OTUs are grouped in a single cluster. The binary relation $\leqslant$ on P_5 is defined such that for all $x, y \in P_5$, $x \leqslant y$ if and only if every cluster of y is a subset of some cluster of x; thus, for example, $\{1234, 5\} \leqslant \{1, 234, 5\}$. Again the reader can verify that $\leqslant$ is a partial order and that $0 = \{12345\}$. I desire such properties to hold in the general case.

(A2) $(X, \leqslant)$ is a poset with 0.

Ordering classifications by structure suggests that amount of classification structure or *complexity* be measured in a manner consistent with the partial order. Such complexity measures have been proposed for use with n-trees. The *component information* [25, 26] of an n-tree counts the number of its *informative components* (i.e., clusters not also present in 0):

$$\mathrm{cmp}(x) = |x| - |0|.$$

The *total information* [25, 26] of an n-tree counts the occurrences of OTUs in the informative components:

$$\mathrm{ttl}(x) = \sum_{A \in x} |A| - \sum_{A \in 0} |A|.$$

The *term information* [26] of an n-tree counts in each informative component all except one of the OTUs:

$$\mathrm{trm}(x) = \mathrm{ttl}(x) - \mathrm{cmp}(x).$$

The *levels sum* [35] of an n-tree counts the number of times that distinct OTUs occur together in informative components:

$$\mathrm{lvl}(x) = \sum_{A \in x} \binom{|A|}{2} - \sum_{A \in 0} \binom{|A|}{2}.$$

Component information and levels sum have also been used to measure the complexity of structure of partitions in P_n [3, 6]. These examples suggest that any measure of complexity should be a real-valued function c on X such that: c is *strictly isotone* [i.e., $c(x) < c(y)$ whenever $x < y$]; c measures complexity on an *interval scale* [2] [i.e., if $x \leqslant y$, then $c(y) - c(x)$ is a meaningful measure of the difference in the complexities of x and y]; and 0 exhibits no complexity [i.e., $c(0) = 0$].

(A3) c is a strictly isotone real-valued function measuring complexity on an interval scale.

(A4) $c(0) = 0$.

Relationships between classification complexity and more general complexity concepts can be established by giving the poset $(X, \leqslant)$ a graph-theoretic formulation. An undirected *graph* $G = (V, E)$ consists of a finite nonempty set V of *vertices* together with a prescribed set E of unordered pairs of distinct vertices of V. Each pair $e = \{x, y\}$ in E is an *edge* that *joins* vertices x and y. A *path* P_{xy} in G between vertices x and y is a sequence $x = x_0, x_1, \ldots, x_m = y$ such that $\{x_i, x_{i+1}\}$ is an edge in E for $0 \leqslant i < m$. G is

connected if a path in G exists between each pair of distinct vertices in V. G is *weighted* if a real-valued function w on E associates with each edge e a weight $w(e)$. The weight of a graph $G = (V, E)$ is simply the sum of its edge weights: $w(G) = \sum_{e \in E} w(e)$. Readers can consult Harary [16] or any standard graph-theory text for elaborations of such concepts, but be forewarned that graph-theoretic terminology is not standardized among texts.

Each poset $(X, \leqslant)$ has associated with it a graph that exhibits the essential structure of the partial order. For $x, y \in X$ let y *cover* x if $x < y$, and $x < z \leqslant y$ implies that $y = z$. The *covering graph* $\mathrm{CG}(X)$ is the graph with vertex set X and with an edge joining x and y if and only if x covers y or y covers x in the poset X. When a real-valued function c is defined on the poset X [a situation denoted by $(X, \leqslant, c)$], then $\mathrm{CG}(X)$ is weighted by requiring for each edge $e = \{x, y\}$ in $\mathrm{CG}(X)$ that $w(e) = |c(x) - c(y)|$.

Classification complexity in X can be interpreted in terms of most parsimonious paths in the weighted covering graph $\mathrm{CG}(X)$. To see this, let the function $\delta_c : X^2 \to \mathbb{R}^{\geqslant 0}$ be defined such that for $x, y \in X$

$$\delta_c(x, y) = \begin{cases} 0 & \text{if } x = y, \\ \text{minimum weight of all paths in} & \\ \quad \mathrm{CG}(X) \text{ between } x \text{ and } y & \text{otherwise.} \end{cases}$$

LEMMA 1

If the poset $(X, \leqslant, c)$ *satisfies* (A1)–(A4), *then* $c(x) = \delta_c(0, x)$ *for every* $x \in X$.

Proof. $\mathrm{CG}(X)$ exists and is connected because of (A1) and (A2), and so δ_c is well defined. Since c is strictly isotone by (A3), every edge in $\mathrm{CG}(X)$ has positive weight, so that each path $P_{0x} : 0 = x_0, x_1, \ldots, x_m = x$ of minimum path weight has $x_i \leqslant x_{i+1}$ for $0 \leqslant i < m$. But then $\delta_c(0, x) = w(P_{0x}) = \sum_{i=0}^{m-1} |c(x_{i+1}) - c(x_i)| = \sum_{i=1}^{m} c(x_i) - \sum_{i=0}^{m-1} c(x_i) = c(x) - c(0) = c(x)$, the last step following by (A4). ∎

In view of Lemma 1, one can visualize the complexity $c(x)$ of classification x as representing the length of a most parsimonious description of x. When y covers x in X, the weight $|c(y) - c(x)|$ of edge $\{x, y\}$ in $\mathrm{CG}(X)$ is the length of a parsimonious description of how classification y may be constructed from classification x; thus $c(x)$ is the length of a shortest description in $(X, \leqslant)$ of how x is constructed from the trivial classification 0. This interpretation of classification complexity is consistent with Papentin's general concept [28, p. 421] that "the length of the shortest possible description... of a pattern in a given language L is used as a measure of its complexity."

In view of Lemma 1, it seems appropriate to investigate the properties exhibited by δ_c when (A1)–(A4) are satisfied. The reader can verify that in

these conditions δ_c satisfies the properties

(M1) For every $x \in X$, $\delta_c(x, x) = 0$;
(M2) For all $x, y \in X$, $\delta_c(x, y) = 0$ implies $x = y$;
(M3) For all $x, y \in X$, $\delta_c(x, y) = \delta_c(y, x)$;
(M4) For all $x, y, z \in X$, $\delta_c(x, y) \leqslant \delta_c(x, z) + \delta_c(z, y)$;

and so is a *metric* measuring the *minimum weighted path distance* [24] between classifications. Furthermore δ_c clearly satisfies the betweenness property

(M5) For all $x, y, z \in X$ with $x \leqslant z \leqslant y$,

$$\delta_c(x, y) = \delta_c(x, z) + \delta_c(z, y).$$

The behavior of δ_c is less obvious when betweenness is defined in the following way. For each subset S of X, let $\mathrm{lb}(S) = \{z \in X : z \leqslant x \text{ for all } x \in S\}$, $c^-(S) = \max\{c(z) : z \in \mathrm{lb}(S)\}$, and $\mathrm{lb}_c(S) = \{z \in \mathrm{lb}(S) : c(z) =$

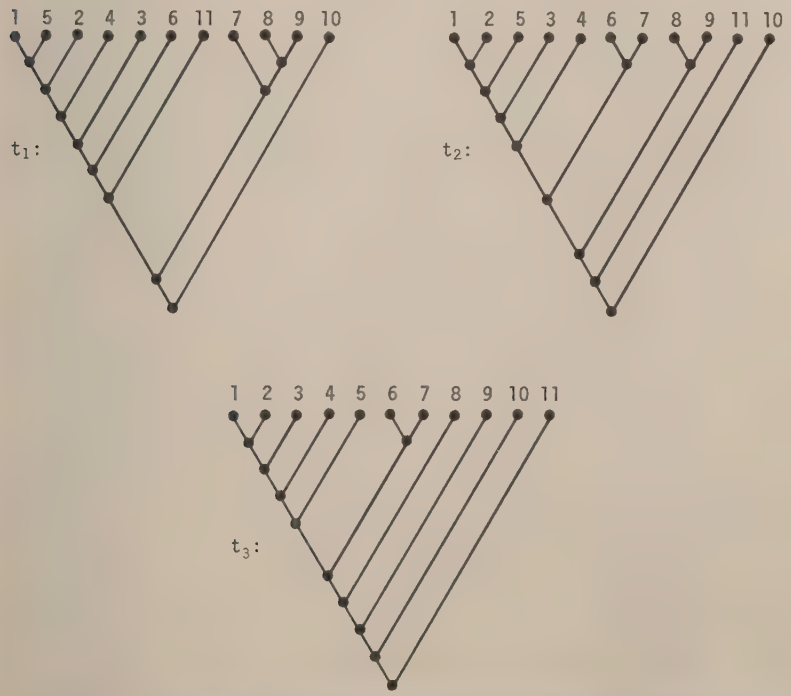

FIG. 2. 11-trees showing relationships among species of *Dipodomys* based on the allozyme data of Johnson and Selander [17]. Modified from Figure 1 of Schnell, Best, and Kennedy [34].

$c^-(S)$); when exactly two classifications are involved, I abuse the notation and write simply $\mathrm{lb}(x, y)$, $c^-(x, y)$, and $\mathrm{lb}_c(x, y)$. Although each classification $z \in \mathrm{lb}_c(x, y)$ lies on a path of weight $\delta_c(x, z) + \delta_c(z, y)$ between x and y, such paths are not necessarily minimum-weight paths between x and y. Hence one can ask for the conditions under which δ_c satisfies the betweenness property

(M6) For all $x, y, z \in X$ with $z \in \mathrm{lb}_c(x, y)$,

$$\delta_c(x, y) = \delta_c(x, z) + \delta_c(z, y).$$

Section 3 includes a characterization of such conditions.

Each classification in $\mathrm{lb}_c(x, y)$ necessarily exhibits a maximum amount of structure shared by both x and y, and so is called a *strict consensus* of x and y. Recently considerable interest has been expressed in consensus methods [5, 18–20, 27, 31, 37]. A consensus method is a function γ that maps each

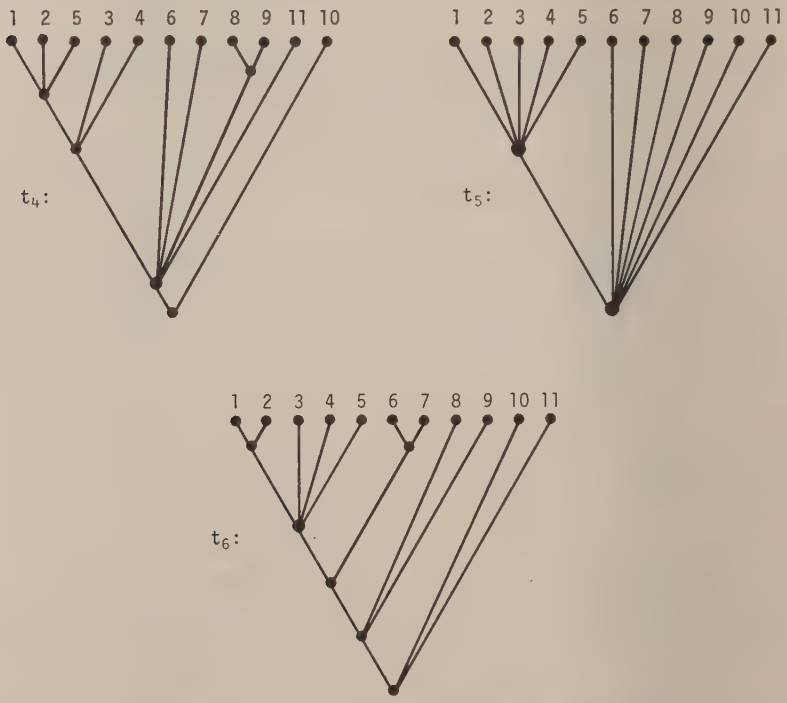

FIG. 3. Consensus classifications for pairs of 11-trees from Figure 2: $t_4 \in \sigma_{\mathrm{cmp}}(\{t_1, t_2\})$, $t_5 \in \sigma_{\mathrm{cmp}}(\{t_1, t_3\})$, and $t_6 \in \sigma_{\mathrm{cmp}}(\{t_2, t_3\})$. The definitions guarantee that $\mu(\{x, y\}) = \sigma_{\mathrm{cmp}}(\{x, y\})$.

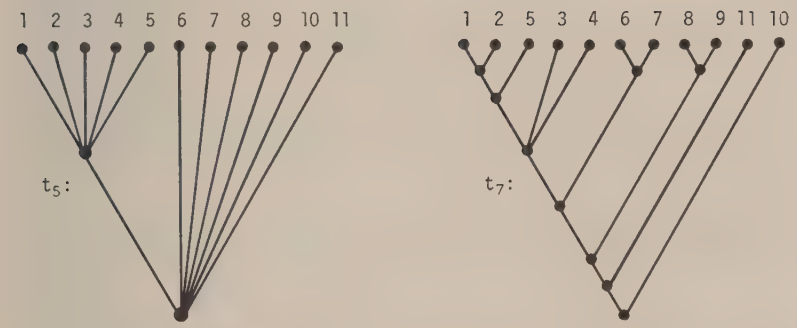

FIG. 4 Consensus classifications for the three 11-trees from Figure 2: $t_5 \in \sigma_{\text{cmp}}(\langle t_1, t_2, t_3 \rangle)$ and $t_7 \in \mu(\langle t_1, t_2, t_3 \rangle)$.

subset $S \subseteq X$ of classifications of interest into a corresponding subset $\gamma(S) \subseteq X$ of consensus classifications. Thus the *strict consensus method* for a poset $(X, \leqslant, c)$ satisfying (A1)–(A4) is simply a function σ_c such that $\sigma_c(S) = \text{lb}_c(S)$ for each subset S of X. Margush and McMorris [18] describe for T_n a quite different consensus method based on the majority rule concept: the *majority-rule consensus method* is a function μ such that for each subset S of T_n, $\mu(S) = \langle M(S) \rangle$, where $M(S) = \{A \subseteq N : A \in x$ for more than half the x's in $S\}$. Figure 3 and Figure 4 illustrate the application of these consensus methods to the 11-trees shown in Figure 2.

3. DISTANCE BETWEEN CLASSIFICATIONS

This section demonstrates fundamental relationships that exist among the concepts of distance between classifications, strict consensus of two classifications, and classification complexity. These relationships follow directly from basic results in poset theory that appear in Monjardet [24].

Minimum weighted path distance $\delta_c(x, y)$ between classifications is an intuitively appealing parsimonious concept, but it is not apparent how one enumerates such paths to compute the actual value of $\delta_c(x, y)$. However, if a minimum weighted path is known to pass through any strict consensus $z \in \sigma_c(x, y)$, then the δ_c computation reduces to possibly simpler computations involving c, since $\delta_c(x, y) = \delta_c(x, z) + \delta_c(z, y) = c(x) + c(y) - 2c^-(x, y)$. Thus it seems appropriate to investigate the circumstances under which δ_c satisfies (M6).

(Q1) Under what conditions does $\delta_c(x, y) = \delta_c(x, z) + \delta_c(z, y)$ for all $x, y, z \in X$ with $z \in \sigma_c(x, y)$?

If such computations are less complex than the more laborious enumeration of minimum weighted paths, then as a practical consideration it may be

appropriate to use them as a measure of dissimilarity between classifications. To do this in any poset $(X, \leqslant, c)$ satisfying (A1)–(A3), let the function $d_c: X^2 \to \mathbb{R}^{\geq 0}$ be defined such that $d_c(x, y) = c(x) + c(y) - 2c^-(x, y)$ for all $x, y \in X$. Clearly d_c satisfies (M1)–(M3), (M5), and (M6); but whether d_c satisfies the *triangle inequality* (M4) is a problem of a somewhat more subtle nature.

(Q2) Under what conditions is d_c a metric?

From the definitions it is clear that $d_c(x, y) \geq \delta_c(x, y)$ for all $x, y \in X$; it is even conceivable that $d_c = \delta_c$ for certain posets $(X, \leqslant, c)$.

(Q3) Under what conditions does $d_c = \delta_c$?

Although these questions are phrased in terms of dissimilarity measures, it is reasonable to expect that their answers involve characteristics of classification complexity c. Observe, for example, that if $c(x) + c(y) \leqslant c^-(x, y) + c(u)$ for all $x, y, u \in X$ with $x \leqslant u$ and $y \leqslant u$, then minimum-weight paths between x and y are forced to pass through classifications in $\sigma_c(x, y)$, so that $d_c = \delta_c$. In fact, in any poset $(X, \leqslant)$ satisfying (A1) and (A2), a real-valued function v on X is called a *lower valuation* if $v(x) + v(y) \leqslant v^-(x, y) + v(u)$ for all $x, y, u \in X$ with u an upper bound of x and y. Thus in answering (Q3) it seems relevant to ask

(Q4) Under what conditions is c a lower valuation?

These questions are answered elegantly by the following result.

THEOREM 2 (Monjardet [24, p. 175])

If poset $(X, \leqslant, c)$ *satisfies* (A1)–(A3), *then the following statements are equivalent*:

(i) *For all* $x, y, z \in X$ *with* $z \in \sigma_c(x, y)$, $\delta_c(x, y) = \delta_c(x, z) + \delta_c(z, y)$.
(ii) d_c *is a metric.*
(iii) $d_c = \delta_c$.
(iv) c *is a lower valuation.*

Any metric d on a poset $(X, \leqslant)$ satisfying (A1) and (A2) is called a (*lower*) *valuation metric* if there exists a strictly isotone lower valuation c such that $d = d_c$. It follows from Theorem 2 that all valuation metrics satisfy the betweenness properties (M5) and (M6). The next result provides a characterization of valuation metrics by establishing that only valuation metrics satisfy (M5) and (M6).

THEOREM 3 (Monjardet [24, p. 176])

Let $(X, \leqslant)$ *be a poset satisfying* (A1) *and* (A2), d *a real-valued function on* X^2, C *a real constant, and* c *a real-valued function on* X *such that* $c(x) =$

$d(0, x) + C$ for all $x \in X$. If d satisfies (M5) *and* (M6), *then $d = d_c$. Moreover, d is a metric if and only if c is a strictly isotone lower valuation.*

In summary, strictly isotone lower valuations are useful as measures of classification complexity, since the corresponding valuation metrics exhibit desirable properties: they are characterized by the betweenness properties (M5) and (M6); and they are minimum-weighted-path metrics in which minimum weighted paths between classifications pass through strict-consensus classifications. Thus I complete the specification of assumptions underlying this comparison model by requiring of c that it be a lower valuation.

(A5) c is a lower valuation.

The complexity measures described for P_n and T_n satisfy (A5). Boorman and Olivier [7] establish that component information (cmp) is a lower valuation on $(P_n, \leqslant)$ and that levels sum (lvl) is a lower valuation for the poset $(P_n, \geqslant)$ in which $\{1,2,\ldots,n\}$ is taken as the zero; Arabie and Boorman [3] and Boorman and Arabie [6] describe the corresponding valuation metrics. On the other hand, $(T_n, \leqslant)$ is a meet semilattice in which the meet of x and y is their set intersection, while if x and y have any upper bound in $(T_n, \leqslant)$, then their join exists and is their set union. From this it follows easily that component information (cmp), total information (ttl), term information (trm) and levels sum (lvl) are all strict isotone lower valuations on $(T_n, \leqslant)$.

4. COMPLEXITY OF SETS OF CLASSIFICATIONS

The comparison model requires that structural complexity $c(x)$ of a classification x be measured by a strictly isotone lower valuation c. Complexity $c(x)$ has several interpretations within the model: it is the weight of a minimum weighted path in $\mathrm{CG}(X)$ between 0 and x; it is the weight, in fact, of a most parsimonious of all weighted subtrees of $\mathrm{CG}(X)$ that include 0 and x as vertices; intuitively it is the length of a shortest description of how x can be constructed from 0. It seems natural to extend the concept of complexity from single classifications to classification sets in such a way as to preserve these interpretations to the extent possible. Let $\mathrm{wmpt}(R)$ denote the weight of a most parsimonious subtree of $\mathrm{CG}(X)$ that includes in its vertex set the subset R of X: $\mathrm{wmpt}(R) = \min\{w(T) : T = (V, E)$ is a subtree of $\mathrm{CG}(X)$, R is a subset of $V\}$. For notational simplicity denote the union $S \cup \{x\}$ by $S + x$; then *set complexity* is a real-valued function c^* defined for arbitrary subsets S of X by requiring that $c^*(S) = \mathrm{wmpt}(S + 0)$.

The concepts of classification complexity c and set complexity c^* coincide when exactly one classification x is involved, since $c^*(\{x\}) = \mathrm{wmpt}(\{0, x\}) = d_c(0, x) = c(x)$. When exactly two classifications x and y are involved,

classification complexity provides an upper bound on set complexity $c^*(\{x, y\})$, by the following argument. If z is any strict consensus of x and y, then $c(z) = c^-(x, y)$ is the weight of a minimum weighted path between 0 and z, while $d_c(x, y)$ is the sum of weights for minimum weighted paths between z and x and between z and y. Thus $c^*(\{x, y\}) \leqslant c^-(x, y) + d_c(x, y) = c(x) + c(y) - c^-(x, y)$; in fact, equality always holds.

LEMMA 4

If the poset $(X, \leqslant, c)$ *satisfies* (A1)–(A5), *then* $c^*(\{x, y\}) = c(x) + c(y) - c^-(x, y)$ *for all* $x, y \in X$.

Proof. The result is immediate if $x \leqslant y$ or $y \leqslant x$. In the remaining case x, y, and 0 are distinct vertices in any minimum-weight subtree T of $CG(X)$ that contains them. If 0 is an internal vertex of T, then minimality of T forces 0 to be a strict consensus of x and y. If 0 is an endpoint of T, then T contains a single vertex z of degree 3, and minimality of T forces z to be a strict consensus of x and y. Either way, $c^*(\{x, y\}) = \text{wmpt}(\{0, x, y\}) = w(T) = c^-(x, y) + d_c(x, y) = c(x) + c(y) - c^-(x, y)$. ∎

Figure 5 exhibits a parsimonious tree in the cover graph of $(T_{11}, \leqslant, \text{cmp})$ for $S + 0$, where S comprises the three 11-trees in Figure 2. Since I have no proof that the tree is most parsimonious, it follows only that $c^*(S) \leqslant 21$.

For the general case when S contains more than two classifications, the computational problem appears to be quite difficult. To see this, restate the problem as follows.

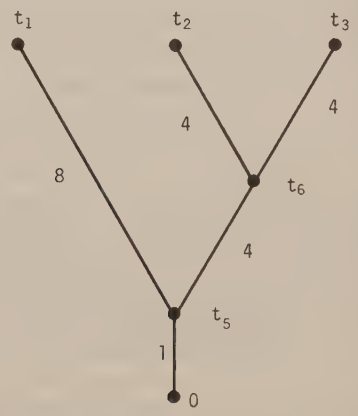

FIG. 5 A parsimonious tree in $CG(T_{11})$ for the set $S + 0$, where $S = \langle t_1, t_2, t_3 \rangle$. Complexity is measured by cmp.

SET COMPLEXITY PROBLEM (SCP)

Instance: Poset $(X, \leqslant, c)$ satisfying (A1)–(A5), cover graph CG(X) in which each edge $e = \langle x, y \rangle$ has weight $w(e) = |c(x) - c(y)|$, a subset $R + 0$ of X, and a positive integer bound B.

Question: Is there a subtree of CG(X) that includes all of the vertices of $R + 0$ and such that the sum of the weights of the edges in the subtree is no more than B?

SCP is a *decision problem*, since each SCP solution is either "yes" or "no." SCP belongs to a special class of decision problems that are called *NP-complete* (Garey and Johnson [14] give a thorough presentation of NP-completeness.)[1] NP-complete problems are all equally difficult in the sense that an efficient algorithm solving any one NP-complete problem would enable *every* NP-complete problem to be solved efficiently. In addition, NP-complete problems are so difficult that no one has yet succeeded in designing an efficient algorithm to solve any one of the many known NP-complete problems. Thus it is unlikely that an efficient algorithm can be designed to solve SCP, and so it is desirable to design suitable approximation procedures to solve it. As Graham and Foulds [15] observe: "... the fact that a problem is NP-complete is considered justification for heuristic procedures to be applied to it, that is, procedures which do not guarantee to produce an optimal solution for every instance of the problem. The challenge is to find heuristics with good performance guarantees which [also have efficient algorithms]." This challenge must be met if set complexity is to be used in a practical context.

5. MEASURING CONSENSUS FIT

Since a consensus z of S represents structure that is judged common, by some criterion, to the classifications in S, it seems appropriate to measure the extent to which z actually fits S. Two complementary factors contribute to such a measurement: the extent to which z represents the classifications in S, and the extent to which z fails to represent them. The problem is related to Papentin's differentiation of pattern complexity into organized and unorganized components [28, p. 424]:

> By organized complexity we understand the length of the minimal description of the rules or laws underlying the formation of a pattern, and by unorganized complexity the length of the minimal description of the random aspects of the pattern, such as... the description of

[1] Since Graham and Foulds [15] prove that the Steiner problem for the n-cube (SPQ) is NP-complete, the NP-completeness of SCP follows by transforming SPQ into it. SPQ involves a subset of R vertices. SPQ (involving R) transforms to SPQ (involving $R + 0$); it in turn transforms to SCP (involving $R + 0$).

pattern aspects which cannot be described by rules. In many cases total complexity will simply be the sum of unorganized and organized complexities.

Figure 6 illustrates the problem. The consensus classification z seems a reasonably close fit to S_1, since its complexity $c(z)$ is large relative to the magnitude $\text{wmpt}(S_1 + z)$ of a most parsimonious description of the set $S_1 + z$; the fit of z to S_2 seems less satisfactory, since $c(z)$ is small relative to $\text{wmpt}(S_2 + z)$. In either case, $c(z)$ presumably measures structure common to the classifications in S_i and so estimates the organized complexity of S_i, while $\text{wmpt}(S_i + z)$ in its fashion estimates the unorganized complexity of S_i. The sum of these terms is always an overestimate of the complexity of S_i.

LEMMA 5

If the poset $(X, \leqslant, c)$ *satisfies* (A1)–(A5), *then* $c^*(S) \leqslant c(z) + \text{wmpt}(S + z)$ *for all subsets* S *of* X *and* $z \in X$.

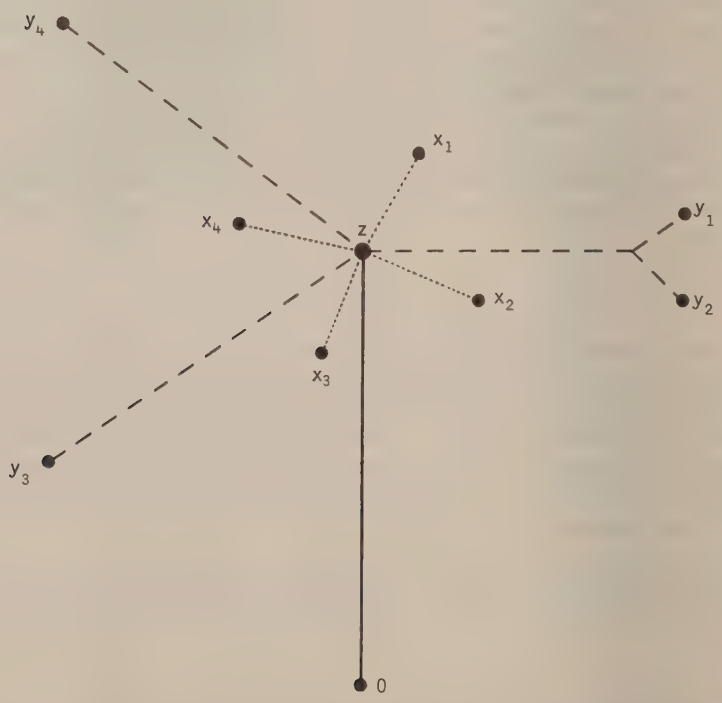

FIG. 6 Measuring consensus fit. z is the consensus of both $S_1 = \{x_1, \ldots, x_4\}$ and $S_2 = \{y_1, \ldots, y_4\}$; yet z seems a closer fit to the classifications in S_1 than to the classifications in S_2.

Proof. Let T be a most parsimonious tree that contains the vertex set $S+z$. Let tree T' be obtained from T by adding edges to obtain a most parsimonious path between 0 and z. Then $c^*(S) \leqslant w(T') \leqslant w(T)+c(z) = \text{wmpt}(S+z)+c(z)$. ∎

If meaningful comparisons of fit are to be made between sets of different size, then the estimate of unorganized complexity of S must be averaged over the classifications in S. With this done, the quantity $c(z)+[\text{wmpt}(S+z)/|S|]$ estimates the average total complexity of classifications in S relative to the consensus z. It can be used to define a real-valued measure α of the fit of z to S by requiring that

$$\alpha(z,S) = \begin{cases} 0 & \text{if } S=\varnothing \text{ or } z=0, \\ \dfrac{1}{1+\dfrac{\text{wmpt}(S+z)}{c(z)\cdot|S|}} & \text{otherwise.} \end{cases}$$

As a measure of fit, α exhibits reasonable properties. Evidently $\alpha(z,S)$ measures, for classifications in S relative to z, the ratio of average organized to average total classification complexity. The value of α is normalized so that $0 \leqslant \alpha(z,S) \leqslant 1$; the extreme values of α arise in nontrivial cases, since $\alpha(0,S)=0$ when $S \neq \varnothing$, and $\alpha(z,\{z\})=1$ when $z \neq 0$. In nontrivial cases z is interpreted as a relatively good (poor) fit to the classifications in S if $\alpha(z,S)$ is relatively close to 1 (0). Since the size effect of S has been eliminated from the estimate of unorganized complexity, meaningful comparisons can be made between $\alpha(z_i,S_i)$ and $\alpha(z_j,S_j)$ when $|S_i| \neq |S_j|$.

To illustrate the computation and use of α, consider the fit of the majority-rule consensus t_7 (in Figure 4) to the set $S=\{t_1,t_2,t_3\}$ of 11-trees (in Figure 2) when complexity is measured by component information (cmp). Since $\text{wmpt}(S+t_7)=17$ (see Figure 7), $\text{cmp}(t_7)=8$, and $|S|=3$,

$$\alpha(t_7,S) = \frac{1}{1+\dfrac{17}{8\times 3}} = 0.59.$$

By comparison, consider the fit of the strict consensus t_5 (in Figure 4) to S; since $18 \leqslant \text{wmpt}(S+t_5) \leqslant 20$ (see Figure 7) and $\text{cmp}(t_5)=1$, it follows that

$$0.13 = \frac{1}{1+\dfrac{20}{1\times 3}} \leqslant \alpha(t_5,S) \leqslant \frac{1}{1+\dfrac{18}{1\times 3}} = 0.14.$$

These results are consistent with the behavior typically exhibited by the strict

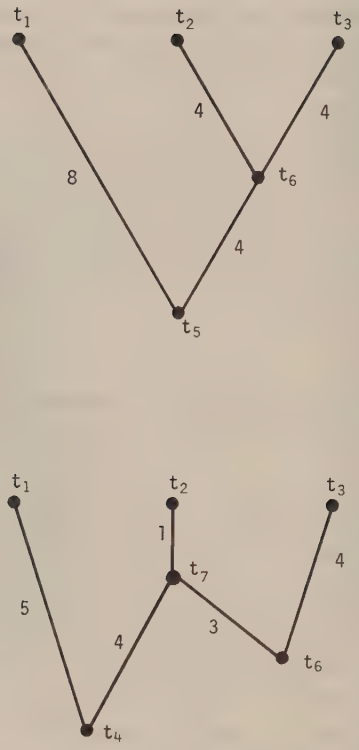

FIG. 7 Parsimonious trees in $CG(T_{11})$ for the sets $S + t_5$ and $S + t_7$, where $S = \langle t_1, t_2, t_3 \rangle$. Complexity is measured by cmp. One can show that $18 \leqslant \text{wmpt}(S + t_5) \leqslant 20$ and $\text{wmpt}(S + t_7) = 17$.

consensus method: since the structure in a strict consensus z must also be exhibited by every classification in S, $c(z)$ is typically small, $\text{wmpt}(S + z)$ is typically large, and consequently $\alpha(z, S)$ is typically small. The example also demonstrates that α may be useful for comparisons even when exact values of $\text{wmpt}(S + z)$ are unknown.

The definition of α indicates that an efficient algorithm to compute $\alpha(z, S)$ could be constructed from efficient algorithms to compute $c(z)$ and $\text{wmpt}(S + z)$. Of these subproblems the computation of $\text{wmpt}(S + z)$ is more difficult; when formulated as a decision problem it, like SCP, is known to be NP-complete [9]. As with SCP, it is unlikely that an efficient algorithm can be designed to compute exact values of $\text{wmpt}(S + z)$ and thus $\alpha(z, S)$.

Work remains to be done concerning computation and interpretation of this measure α of consensus fit. We need to design and implement an efficient algorithm that obtains a close approximation to $\text{wmpt}(R)$ for

subsets R of X. We need to provide statistical information concerning the distribution of α values so that users may determine the significance of computed $\alpha(z, S)$ values. My colleagues[2] and I hope to contribute to the solutions of these problems.

REFERENCES

1 E. N. Adams III, Consensus techniques and the comparison of taxonomic trees, *Syst. Zool.* 21:390–397 (1972).

2 M. R. Anderberg, *Cluster Analysis for Applications*, Academic, New York, 1973, p. 27.

3 P. Arabie and S. A. Boorman, Multidimensional scaling of measures of distance between partitions, *J. Math. Psych.* 10:148–203 (1973).

4 J. Archie, Definition, criteria and testing of the predictive value of classifications, Ph.D. dissertation, State Univ. of New York at Stony Brook, 1980.

5 J. P. Barthelemy and B. Monjardet, The median procedure in cluster analysis and social choice theory, *Math. Social Sci.* 1:235–267 (1981).

6 S. A. Boorman and P. Arabie, Structural measures and the method of sorting, in *Multidimensional Scaling, Volume I: Theory* (R. N. Shepard, A. K. Romney and S. B. Nerlove, Eds.), Seminar Press, New York, 1972, pp. 225–249.

7 S. A. Boorman and D. C. Olivier, Metrics on spaces of finite trees, *J. Math. Psych.* 10:26–59 (1973).

8 W. H. E. Day, The complexity of computing metric distances between partitions, *Math. Social Sci.* 1:269–287 (1981).

9 W. H. E. Day, Computationally difficult parsimony problems in phylogenetic systematics, *J. Theoret. Biol.*, 103:429–438 (1983).

10 A. J. Dobson, Comparing the shapes of trees, in *Lecture Notes in Mathematics*, Vol. 452 (A. P. Street and W. D. Wallis, Eds.), Springer, Heidelberg, 1975, pp. 95–100.

11 J. S. Farris, On the phenetic approach to vertebrate classification, in *Major Patterns in Vertebrate Evolution* (M. K. Hecht, P. C. Goody, and B. M. Hecht, Eds.), NATO Advanced Study Institute Series, Vol. 14, Plenum, New York, 1977, pp. 823–850.

12 J. S. Farris, On the naturalness of phylogenetic classification, *Syst. Zool.* 28:200–214 (1979).

13 J. S. Farris, The information content of the phylogenetic system, *Syst. Zool.* 28:483–519 (1979).

14 M. R. Garey and D. S. Johnson, *Computers and Intractability. A Guide to the Theory of NP-Completeness*, Freeman, San Francisco, 1979.

15 R. L. Graham and L. R. Foulds, Unlikelihood that minimal phylogenies for a realistic biological study can be constructed in reasonable computational time, *Math. Biosci.* 60:133–142 (1982).

16 F. Harary, *Graph Theory*, Addison-Wesley, Reading, Mass., 1969.

17 W. E. Johnson and R. K. Selander, Protein variation and systematics in kangaroo rats (genus *Dipodomys*), *Syst. Zool.* 20:377–405 (1971).

18 T. Margush and F. R. McMorris, Consensus n-trees, *Bull. Math. Biol.* 43:239–244 (1981).

[2] N. F. Rehner, On the approximate computation of a consensus index, annual North American meeting of the Classification Society, Philadelphia, 29–31 May 1983.

19 F. R. McMorris, D. B. Meronk, and D. A. Neumann, A view of some consensus methods for trees, in *Numerical Taxonomy*: *Proceedings of NATO Advanced Study Institute* (J. Felsenstein, Ed.), Springer, 1983.

20 F. R. McMorris and D. A. Newmann, Consensus functions defined on trees, in *Mathematical Social Sciences*, 4:131–136 (1983).

21 M. F. Mickevich, Taxonomic congruence, *Syst. Zool.* 27:143–158 (1978).

22 M. F. Mickevich, Taxonomic congruence: Rohlf and Sokal's misunderstanding, *Syst. Zool.* 29:162–176 (1980).

23 M. F. Mickevich and J. S. Farris, The implications of congruence in *Menidia*, *Syst. Zool.* 30:351–370 (1981).

24 B. Monjardet, Metrics on partially ordered sets—a survey, *Discrete Math.* 35:173–184 (1981).

25 G. Nelson, Cladistic analysis and synthesis: Principles and definitions, with a historical note on Adanson's *Familles des Plantes* (1763–1764), *Syst. Zool.* 28:1–21 (1979).

26 G. Nelson and N. Platnick, *Systematics and Biogeography. Cladistics and Vicariance*, Columbia U. P., New York, 1981, pp. 255–257.

27 D. A. Neumann, Faithful consensus methods for *n*-trees, *Math. Biosci.*, 63:271–287 (1983).

28 F. Papentin, On order and complexity. I. General considerations, *J. Theoret. Biol.* 87:421–456 (1980).

29 F. Papentin, On order and complexity. II. Application to chemical and biochemical structures, *J. Theoret. Biol.* 95:225–245 (1982).

30 F. J. Rohlf, Methods of comparing classifications, *Annual Rev. Ecol. Syst.* 5:101–113 (1974).

31 F. J. Rohlf, Consensus indices for comparing classifications, *Math. Biosci.* 59:131–144 (1982).

32 F. J. Rohlf and R. R. Sokal, Comments on taxonomic congruence, *Syst. Zool.* 29:97–101 (1980).

33 F. J. Rohlf and R. R. Sokal, Comparing numerical taxonomic studies, *Syst. Zool.* 30:459–490 (1981).

34 G. D. Schnell, T. L. Best, and M. L. Kennedy, Interspecific morphologic variation in kangaroo rats (*Dipodomys*): Degree of concordance with genic variation, *Syst. Zool.* 27:34–48 (1978).

35 R. T. Schuh and J. S. Farris, Methods for investigating taxonomic congruence and their application to the Leptopodomorpha, *Syst. Zool.* 30:331–351 (1981).

36 R. T. Schuh and J. T. Polhemus, Analysis of taxonomic congruence among morphological, ecological, and biogeographic data sets for the Leptopodomorpha (Hemiptera), *Syst. Zool.* 29:1–26 (1980).

37 K. T. Shao and F. J. Rohlf, Sampling distribution of consensus indices when all bifurcating trees are equally likely, in *Numerical Taxonomy*: *Proceedings of NATO Advanced Study Institute* (J. Felsenstein, Ed.), Springer, 1983.

38 R. R. Sokal and F. J. Rohlf, The comparison of dendrograms by objective methods, *Taxon* 11:33–40 (1962).

39 R. R. Sokal and F. J. Rohlf, Taxonomic congruence in the Leptopodomorpha re-examined, *Syst. Zool.* 30:309–325 (1981).

40 G. F. Estabrook and F. R. McMorris, When is one estimate of evolutionary relationships a refinement of another? *J. Math. Biol.* 10:367–373 (1980).

Necessary and Sufficient Conditions for a Phenotype System to Have a Factor-Union Representation

GEORGE MARKOWSKY
Computer Sciences Department, IBM T. J. Watson Research Center, Yorktown Heights, New York 10598

Received 8 June 1982; revised 22 April 1983

ABSTRACT

This paper presents an algorithm for determining whether a phenotype system has a factor-union representation, i.e., whether sets of properties (factors) can be assigned to alleles so that the phenotype of each genotype is determined by the union of the factor sets of the alleles. If such a representation is possible, the algorithm can construct one using p or fewer factors, where p is the number of phenotypes. Refinements to the algorithm are presented that often construct representations using few factors. An example is presented to show that some phenotype systems can be represented only by using almost as many factors as there are phenotypes in the system. Finally, it is shown that the problem of finding factor-union representations for multilocus systems reduces to the problem for 1-locus systems.

1. INTRODUCTION

Cotterman [1] investigated the existence of factor-union representations for phenotype systems and derived some partial answers. This paper completely solves the problem and presents an algorithm which always determines whether a phenotype system has a factor-union representation. This algorithm is efficient enough to be used manually on some large examples.

If a factor-union representation is possible for a phenotype system, the algorithm can construct one using the phenotypes themselves as factors. In general, this may use many more factors than necessary, since it is possible to represent 2^k distinct phenotypes with k factors. Refinements to the algorithm can significantly reduce the number of factors required. As an example, two representations are constructed for an example left unresolved in [1]. The first one uses 12 factors, while the second one uses only 4.

An example in this paper shows that some phenotype systems require almost as many factors to represent them as they have phenotypes. This

©Elsevier Science Publishing Co., Inc., 1983
52 Vanderbilt Ave., New York, NY 10017 0025-5564/83/$03.00

example is also of interest because the number of factors required grows quadratically in the number of alleles, which shows that simple representations are not always possible.

The representations constructed by the algorithm need not be based on actual physical factors. Determining a physical basis for a representation is a problem that cannot be solved purely mathematically, but the mathematical results can guide the search for suitable physical factors. Of course, if the algorithm shows that no representation is possible, much valuable time can be saved.

In Section 5, techniques are developed to translate multilocus systems into 1-locus systems so that finding a factor-union representation for one system finds it for the other. Thus, this algorithm can be used with any phenotype system.

2. DEFINITIONS AND NOTATION

Listed below are several simple definitions and concepts required for understanding this paper. Many of them come from [1], which should be consulted for general background on this topic. A few come from the theory of partial orders.

(1) Lowercase letters denote alleles, Greek letters denote genotypes, and uppercase letters denote phenotypes. Pairs of lowercase letters will be used when it is desirable to show the alleles contained in a genotype. Bracketed expressions such as $\langle ab, cd \rangle$ are used to show the genotypes in a phenotype.

(2) If the two alleles in a genotype are the same, it is called a *homozygote*. If they are different, it is called a *heterozygote*.

(3) The symbol $+$ denotes the union of individual sets, Σ the union of a family of sets, and $\prod$ the intersection of a family of sets. The symbols $\leqslant$ and $<$ denote set inclusion and proper set inclusion respectively.

(4) A *factor-union representation* of a phenotype system Φ consists of a set X of factors, called the *factor set*, and an assignment f of subsets of X to the alleles of Φ with the following properties:

(a) If ab and cd are two genotypes in the same phenotype,

$$f(a)+f(b)=f(c)+f(d).$$

The set $f(a)$ is called the *factor set* of a. If ab is in P then

$$f(a)+f(b)=\sum_{x \text{ in } P} f(x).$$

The quantity on the right side is abbreviated $f(P)$ and called the *factor set* of P. It is the union of the factor sets of all alleles in P.

(b) If P and Q are distinct phenotypes, $f(P) \neq f(Q)$. This condition states that phenotypes are determined by their factors.

(5) A *partially ordered set* (*poset*) is a set X together with a partial order $<$ defined on its elements. A *partial order* is any binary relation that satisfies the following two rules:

(a) *Reflexive property*. For all x in X, $x < x$ never holds.

(b) *Transitive property*. For all x, y, z in X, if $x < y$ and $y < z$, then $x < z$.

The string $x \leqslant y$ means that either $x < y$ or $x = y$. By convention, $y > x$ and $y \geqslant x$ mean the same as $x < y$ and $x \leqslant y$. The same symbols are used for partial order and set inclusion. This should cause no confusion, since set inclusion is a form of partial order.

(6) Let $(X, <)$ be a poset and S a subset of X. An element z of X is called an *upper bound for* S if it is greater than or equal to every element of S. An element is the sup or *least upper bound* of S if it is an upper bound that is less than or equal to every upper bound of S. In general, a subset of a poset need not have a sup. A poset in which every finite nonempty subset has a sup is called a *semilattice*.

(7) If X is a set, Power(X) denotes the *power set* of X, which is the set of all subsets of X. Set inclusion is a partial order on Power(X). With this order, Power(x) is a semilattice in which the sup of a family of sets is its union.

3. THE ALGORITHM

The algorithm attempts to construct a partial order on the set of phenotypes using necessary inclusion relations among the phenotypes. If the algorithm fails to construct a partial order, Theorem 2 shows that two distinct phenotypes will have identical factor sets in any factor-union representation. Thus no factor-union representation is possible. On the other hand, if the algorithm constructs a partial order, a factor-union representation can be constructed using Theorem 1.

The following describes the part of the algorithm that determines whether a factor-union system is possible. Theorem 2 proves that the algorithm is correct and shows how to construct a factor-union representation.

FACTOR-UNION REPRESENTATION ALGORITHM (*Existence Part*)

1. Represent each phenotype of Φ by a point.

2. For each ordered pair of distinct phenotypes (P_i, P_j), draw an arrow from P_i to P_j if and only if some genotype contained in P_i has all its alleles

belonging to some genotype in P_j. This step creates a directed graph having the phenotypes as its points and arrows as its edges. This graph is called Graph(Φ). If there is an arrow from P_i to P_j, then in any factor-union representation, the factor set of P_i is a subset of the factor set of P_j.

3. If Graph(Φ) has a *cycle*, i.e., a sequence of arrows from a point to itself, Φ does not have a factor-union representation. A cycle in Graph(Φ) implies that all the factor sets connected by the arrows must be equal.

4. If Graph(Φ) does not have a cycle, define a partial order $<$ on Graph(Φ) as follows: $P_i < P_j$ if and only if there is a sequence of arrows from P_i to P_j. A sequence of arrows from one point to another is called a *path*. [The condition that there are no cycles in Graph(Φ) implies that $<$ is reflexive. Clearly, $<$ is also transitive and so a partial order. If $P_i \leqslant P_j$, the factor set of P_i is a subset of the factor set of P_j in any factor-union representation.]

5. The next stage of the algorithm adds arrows to Graph(Φ) and extends the partial order $<$ until either a cycle appears in Graph(Φ) or there is no further way to carry out the following rule.

ARROW-ADDING RULE

Let *aa* be in P_1, *bb* in P_2, *ab* in P_3, and P_4 a phenotype such that $P_1 \leqslant P_4$ and $P_2 \leqslant P_4$ in Graph(Φ). Then add an arrow from P_3 to P_4 unless there already is such an arrow.

As arrows are added to Graph(Φ), $<$ must be updated. The partial order is always defined in the same way relative to Graph(Φ), i.e., $P < P^*$ if and only if there exists a path in Graph(Φ) from P to P^*.

To use the Arrow-Adding Rule, choose a heterozygote *ab*. Then find the phenotypes P_1 and P_2 that contain the homozygotes *aa* and *bb*. Next, find the upper bounds of P_1 and P_2. Finally, add the arrows from P_3 to the upper bounds. *The algorithm always halts*, since Graph(P) can have at most $p(p-1)/2$ arrows without having a cycle.

Instead of drawing all arrows in Graph(P), I will only draw an arrow from P to Q if $P < Q$ *and* there does not exist an R such that $P < R < Q$. In this case Q is said to *cover* P. This representation, called the *Hasse diagram*, contains all the information of the original graph, but is easier to draw and interpret. The original order can be reconstructed from this reduced order by transitivity. To further simplify the drawings, most arrow heads will be omitted. Line segments will point from lower to higher points.

Example 1 (System Φ_1). This example, from Cotterman [1, p. 8], has 6 alleles and 12 phenotypes. Cotterman was unable to determine whether this system had a factor-union representation.

Figure 1 shows the Hasse diagram of Graph(Φ_1) after the algorithm has finished with stage 2. There are no cycles in it, so that the algorithm proceeds

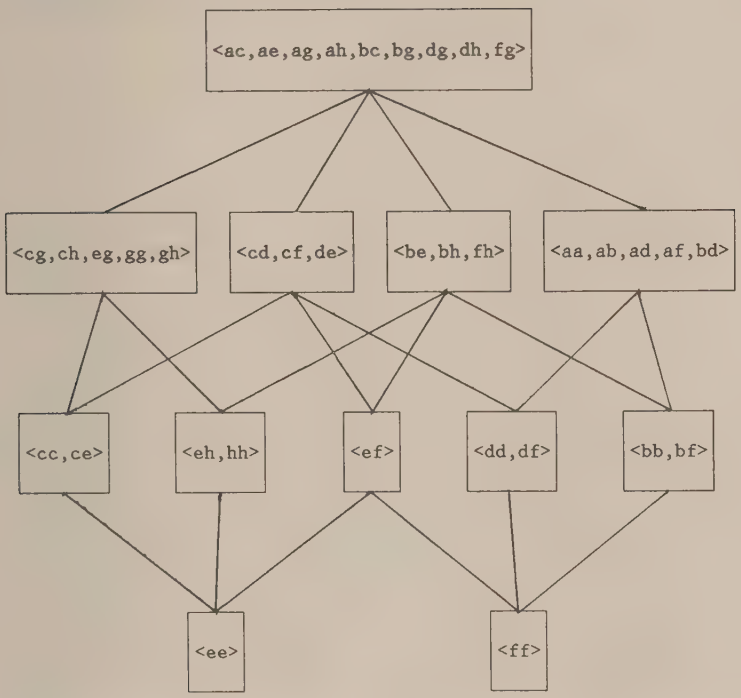

FIG. 1. The Hasse diagram of Graph(Φ_1) after being processed by the algorithm.

to stage 4. Stage 5 lasts a short time, since the Arrow-Adding Rule cannot be applied. Thus Φ_1 has a factor-union representation.

A factor-union representation with 12 factors can be constructed for Φ_1 using Theorem 2. This representation will not be displayed since a representation using only 4 factors is produced in Section 4.

Example 2 (System Φ_2). This example also has 6 alleles. The phenotypes are shown in Figure 2, which displays the results after the first stage of the algorithm is completed. Graph(Φ_2) does not have any cycles, and there are many ways to apply the Arrow-Adding Rule. Theorem 2 shows that the Arrow-Adding Rule can be applied to heterozygotes in any order. If, however, a factor-union representation does not exist, some sequences of heterozygotes will produce a cycle faster than others. For this example, the Arrow-Adding Rule will be applied to the heterozygotes *gx*, *hx*, *hu*, *uy*, and *ty* in order. The intermediate steps are detailed below and the results shown in Figure 3.

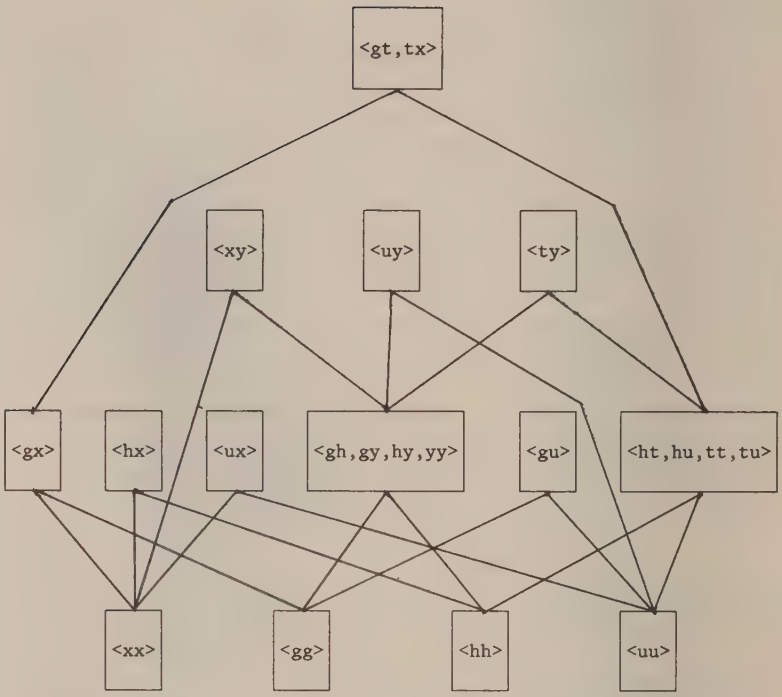

FIG. 2. The Hasse diagram of Graph(Φ_2) after the first stage of the algorithm.

Since $\langle xx \rangle$ and $\langle gg \rangle$ are both less than $\langle xy \rangle$, an arrow must be added from $\langle gx \rangle$ to $\langle xy \rangle$. The arrow from $\langle xx \rangle$ to $\langle xy \rangle$ can now be removed, since there is a path from $\langle xx \rangle$ to $\langle xy \rangle$ through $\langle gx \rangle$.

Since $\langle xx \rangle$ and $\langle hh \rangle$ are both less than $\langle xy \rangle$ and $\langle gt, tx \rangle$, arrows must be added from $\langle hx \rangle$ to $\langle xy \rangle$ and $\langle gt, tx \rangle$.

Since $\langle hh \rangle$ and $\langle uu \rangle$ are both less than $\langle uy \rangle$, add an arrow from $\langle ht, hu, tt, tu \rangle$ to $\langle uy \rangle$. Now the arrow from $\langle uu \rangle$ to $\langle uy \rangle$ can be removed, since there is a path from $\langle uu \rangle$ to $\langle uy \rangle$ through $\langle ht, hu, tt, tu \rangle$.

Again, $\langle ty \rangle$ is greater than $\langle uu \rangle$ and $\langle gh, gy, hy, yy \rangle$, so an arrow is added from $\langle uy \rangle$ to $\langle ty \rangle$. Since $\langle uy \rangle$ is greater than $\langle ht, hu, tt, tu \rangle$ and $\langle gh, gy, hy, yy \rangle$, we add an arrow from $\langle ty \rangle$ to $\langle uy \rangle$. This arrow combined with the previous one creates a cycle involving the two phenotypes $\langle ty \rangle$ and $\langle uy \rangle$. Thus Φ_2 does not have a factor-union representation.

Based on the results just obtained, there is a direct way to show that Φ_2 cannot have a factor-union representation. Suppose, to the contrary, that it did have a factor-union representation f. Then the factors associated with

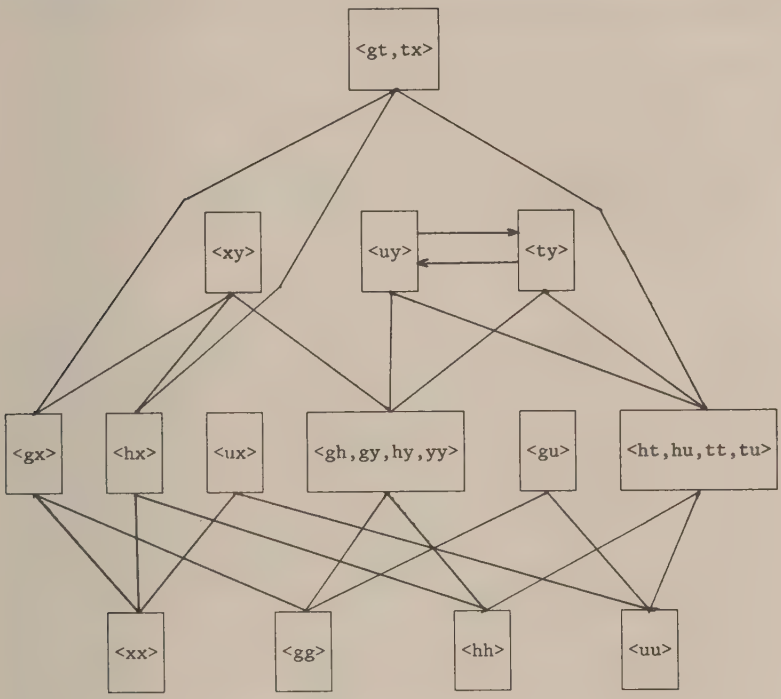

FIG. 3. The Hasse diagram of Graph(Φ_2) when a cycle is first detected.

$\langle uy \rangle$ would be $f(u)+f(y)$. Similarly, the factors associated with $\langle ty \rangle$ would be $f(t)+f(y)$. Since yy is in the same phenotype as gh, it would follow that $f(y)=f(y)+f(y)=f(g)+f(h)$. Similarly, since tt is in the same phenotype as hu, it follows that $f(t)=f(h)+f(u)$. Thus $f(t)+f(y)=f(h)+f(u)+f(g)+f(h)=f(u)+f(g)+f(h)=f(u)+f(y)$. This is impossible, since distinct phenotypes must have distinct factor sets associated with them.

Theorem 1 is a version of Theorem 2.1 in Markowsky [1973], and contains the basic facts about the representation of posets needed to prove that the algorithm is correct. Theorem 2 gives the proof that the algorithm is correct. For additional references and a more complete treatment of these ideas, see Markowsky [1980].

THEOREM 1

Let Δ *be a poset and* $\theta: \Delta \to \text{Power}(\Delta)$ *a function given by the following rule*:

$$\theta(x) = \{y \text{ in } \Delta | y \text{ is not } \geqslant x\}$$

The function θ has the following properties:

(1) *For all x and y in Δ,*

$$x \leqslant y \quad \textit{if and only if} \quad \theta(x) \leqslant \theta(y).$$

Thus distinct elements are mapped into distinct elements.

(2) *Let S be a subset of Δ, and assume that S has a sup, s. Then*

$$\theta(s) = \sum_{t \text{ in } S} \theta(t).$$

Proof. If $x \leqslant y$, and t is not $\geqslant x$, then t is not $\geqslant y$, since $\geqslant$ is transitive. Thus $\theta(x)$ is a subset of $\theta(y)$. Suppose that $\theta(x)$ is a subset of $\theta(y)$. Further, suppose y is not $\geqslant x$. Then y must be an element of $\theta(x)$ and hence of $\theta(y)$. But if y belongs to $\theta(y)$, then y is not $\geqslant y$, which is impossible. Thus $y \geqslant x$.

If s is the sup of S, then $s \geqslant t$ for all t in S. From the first part of this theorem, it follows that $\theta(s)$ contains the union of the $\theta(t)$'s, since it contains each individual $\theta(t)$. Now suppose that there exists an x which belongs to $\theta(s)$ but not to any $\theta(t)$ where t is an element of S. Thus x is not $\geqslant s$, but $x \geqslant t$ for all t in S. Since s is the sup of S, it follows that $x \geqslant s$. This contradicts the fact that x is not in $\theta(s)$. ■

THEOREM 2

If the algorithm described above halts without finding a cycle, Φ has a factor-union representation. If a cycle is found, Φ cannot have a factor-union representation. Furthermore, the heterozygotes in stage 5 of the algorithm can be chosen arbitrarily.

Proof. If Φ has a factor-union representation f, it is impossible for the algorithm to construct a cycle when processing Graph(Φ), since all the arrows introduced by the algorithm correspond to set inclusions that hold in any a factor-union representation.

If a cycle exists in Graph(Φ), at least two distinct phenotypes have the same factor sets, contradicting the assumption that f was a factor-union representation.

If the algorithm halts without finding a cycle, the relation $<$ is a partial order on Graph(Φ). Theorem 1 produces a representation θ assigning subsets of Graph(Φ) to each point in Graph(Φ).

For each allele a in Φ, let $f(a) = \theta(\langle aa \rangle)$. I claim that f is a factor-union representation of Φ. Note that $\theta(\langle ab \rangle) = f(a) + f(b)$, since the Arrow-Adding Rule ensures that $\langle ab \rangle$ is the sup of $\langle aa \rangle$ and $\langle bb \rangle$ in Graph(Φ), and the representation of Theorem 1 transforms sups into unions. If α is a genotype, define $f(\alpha)$ to be the union of the factor sets of the alleles in the genotype. It follows that $f(\alpha) = \theta(P)$, where P is the phenotype containing α.

Since θ maps distinct phenotypes into distinct subsets, f is a factor-union representation of Φ.

The order in which heterozygotes are considered in stage 4 is immaterial, because either a phenotype system has a factor-union representation or it doesn't. If it does, it is impossible for the algorithm to find a cycle, since such a cycle would imply the equality of two distinct phenotypes. Similarly, if a system doesn't have a factor-union representation, the algorithm must eventually discover a cycle, else a factor-union representation could be constructed for the system. ■

4. THE SIZE OF A REPRESENTATION

Example 3 (System Φ_3). This example has $2k$ alleles: $a_1, a_2, \cdots, a_k$ and $x_1, x_2, \cdots, x_k$, and k^2+2k+1 phenotypes:

(1) $\langle a_i a_i \rangle$ for $i = 1,2,\ldots$ (k of this type).
(2) $\langle x_i x_i \rangle$ for $i = 1,2,\ldots$ (k of this type).
(3) $\langle a_i x_j \rangle$ for $i, j = 1,2,\ldots$ (k^2 of this type).
(4) $\langle a_i a_j, x_i x_j | i \neq j$ and $i, j = 1,2,\ldots,k \rangle$ (1 of this type).

I claim that any factor-union representation of Φ_3 must have at least k^2 factors in it. To see this, assume a factor-union representation f for Φ_3 has been given. Let S be factor set of the phenotype $\langle a_1 a_2, \ldots \rangle$, and $S_{i,j}$ the factor set of the phenotype $\langle a_i x_j \rangle$. Here i and j both range over $1,\ldots,k$. Clearly, $S_{i,j}$ is a subset of S for all i and j. Let $C_{i,j}$ be $S - S_{i,j}$. Each $C_{i,j}$ is not empty, since each $S_{i,j}$ must be a proper subset of S.

I claim that the intersection of two distinct $C_{i,j}$'s must be empty. If $(i, j) \neq (i', j')$, then either $i \neq i'$ or $j \neq j'$. In the first case, $S_{i,:} + S_{i',j'}$ contains S, since $S = f(a_i a_{i'})$. The same result holds in the second case. Since the union of $S_{i,j}$ and $S_{i',j'}$ is S, the intersection of $C_{i,j}$ and $C_{i',j'}$ must be empty. Thus the $C_{i,j}$'s are a family of k^2 mutually disjoint, nonempty subsets of S. This implies that S has at least k^2 factors.

Φ_3 has k^2+2k+1 phenotypes. As k increases, the minimum number of factors necessary to represent it asymptotically approaches the number of phenotypes, and is quadratic in the number of alleles.

The following theorem is useful for finding minimal factor-union representations when the poset turns out to be a semilattice. It is an immediate consequence of Proposition 2.6 in [2]. For additional details on this result see [3].

THEOREM 3

Let Δ be a finite semilattice and $\chi: \Delta \to \text{Power}(X)$ a function which is injective (1-*to*-1), *preserves order, and preserves* sups. *Then the number of elements in X is $\geq$ the number of elements of Δ that have exactly one element*

covering them in the Hasse diagram of Δ. *These elements are called meet-irreducibles. Furthermore, if* M *is the set of all meet-irreducible elements in* P, *then* $\theta : \Delta \to \text{Power}(M)$ *given by*

$$\theta(x) = \{y \text{ in } M | y \text{ is not } \geqslant x\}$$

maps distinct elements into distinct elements, preserves order, and preserves sups.

If Graph(Φ) is a semilattice such that in any factor-union representation all sups correspond to unions, Theorem 3 shows how to easily find a factor-union representation using the smallest possible number of factors. Checking whether Graph(Φ) is a semilattice only requires that you check that every pair of elements has an upper bound. Checking whether unions would always correspond to sups is more involved, and is illustrated in the following examples.

It is easy to check that Graph(Φ_1) is a semilattice, with sups being unions. For example, $P = \langle cd, cf, de \rangle$ and $Q = \langle be, bh, fh \rangle$ have $R = \langle ac, ae, \ldots, fg \rangle$ as a sup in the partial order constructed. In any factor-union representation f of Φ_1, $f(R)$ must contain the union of $f(P)$ and $f(Q)$. To apply Theorem 3, it must be the union of $f(P)$ and $f(Q)$. This follows because $f(R) = f(b) + f(c)$, but $f(P) + f(Q) = f(c) + f(d) + f(b) + f(e)$. The argument for all other sups is similar. Thus any factor-union representation of Φ_1 must be a mapping of the type described in Theorem 3.

Graph(Φ_1) has 4 meet-irreducible elements: $\langle aa, \ldots \rangle$, $\langle cg, \ldots \rangle$, $\langle be, \ldots \rangle$ and $\langle cd, \ldots \rangle$. Let these be represented by the symbols w, x, y, and z for brevity. The 4-factor representation of Φ_1 derived from Theorem 3 is illustrated in Figure 4.

The phenotypes corresponding to the meet-irreducible elements have both a label (one of w, x, y, or z) and a factor set of labels. It is important to distinguish between the two. The phenotype $\langle cd, cf, de \rangle$ has the label z, but receives the factor set $\{w, x, y\}$ because the other three meet-irreducible elements, w, x, and y, are not greater than or equal to it. The phenotypes that are not meet-irreducibles only have a factor set associated with them, so there is no confusion.

Example 3 is analyzed similarly. The meet-irreducible elements turn out to be the $\langle a_i x_j \rangle$'s where $i, j = 1, 2, \ldots, k$. Here Theorem 3 shows that k^2 is a possible number of factors for a factor-union representation and that it is also the smallest possible number.

5. MULTILOCUS SYSTEMS

Finding factor-union representations for multilocus systems reduces in a straightforward way to finding factor-union representations for 1-locus sys-

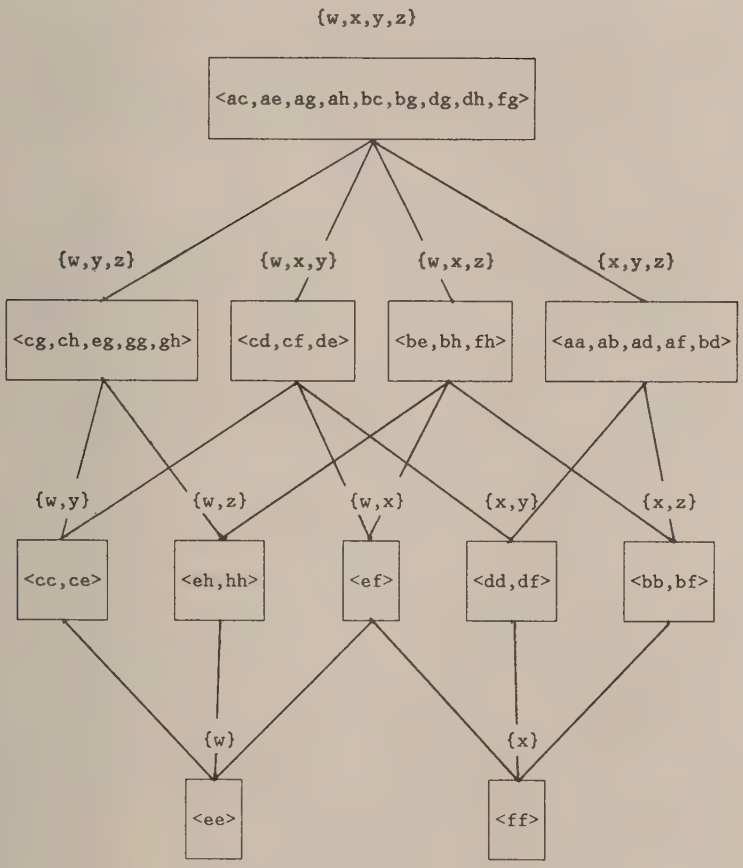

FIG. 4. A 4-factor representation of Φ_1.

tems. A computer is probably required for determining whether even small multilocus systems have a factor-union representation.

DEFINITIONS AND NOTATION

(1) The conventions introduced earlier for alleles and phenotypes are still in effect. A new notation, however, is required to represent genotypes. Thus *aa-bc-dd* represents an individual homozygous at the first and third loci and heterozygous at the second locus. For simplicity, it will be assumed that an allele can occur only at a particular locus. Because of this assumption, distinct genotypes must have different sets of alleles.

(2) The genotype *ab* can be considered the union of *aa* and *bb*. Similarly,

ab-cd-ee can be considered the union of *aa-cd-ee* and *bb-cd-ee*. In general, the union of two or more genotypes is possible if and only if there are no more than two different alleles at each locus.

(3) In the remainder of this paper, the term *homozygote* will refer to a genotype which is homozygous at each locus. The term *heterozygote* will refer to a genotype which is heterozygous at some locus.

(4) Given a multilocus system Φ, a 1-locus system, called the *1-locus model of* Φ and denoted by Trans(Φ), can be derived from it as follows:
   (a) The alleles of Trans(Φ) are derived from the homozygotes of Φ, i.e., for each homozygote of Φ, use a new symbol to represent an allele in Trans(Φ). If α is a homozygote in Φ, let $A(\alpha)$ represent the corresponding allele of Trans(Φ).
   (b) Since each genotype α in Φ can be written as the union of two homozygotes and the union of any two homozygotes is possible, every genotype in Φ corresponds to some genotype in Trans(Φ). Any such genotype is denoted by Trans(α).
   (c) Define the phenotypes as follows. Trans(α) and Trans(β) are in the same phenotype if and only if α and β are in the same phenotype in Φ. For each phenotype P in Φ, Trans(P) denotes the corresponding phenotype in Trans(Φ). [In general, Trans(Φ) has more genotypes than Φ, but the same number of phenotypes.]

(5) A *factor-union representation* for a multilocus system Φ consists of a set of factors and an assignment of subsets of factors to the alleles of Φ, such that the factor sets associated with all the genotypes within a given phenotype are equal, and distinct phenotypes have distinct factor sets. The factor set of a genotype is computed by taking the union of factor sets associated with the alleles at each locus. This definition is essentially the same as the definition for 1-locus systems.

Example 4. Let Φ be a 2-locus system with two alleles at each locus. Suppose the alleles at the first locus are w and x, while the alleles at the second locus are y and z. Φ has nine genotypes: four homozygotes and five heterozygotes. Further, suppose each genotype of Φ forms a distinct phenotype. To construct Trans(Φ), pick four new symbols corresponding to the homozygotes of Φ. I will use the following correspondence:

$$\langle ww\text{-}yy\rangle \to a, \quad \langle ww\text{-}zz\rangle \to b, \quad \langle xx\text{-}yy\rangle \to c, \quad \langle xx\text{-}zz\rangle \to d.$$

Trans(Φ) has ten genotypes and nine phenotypes. Each phenotype, except for $\langle ad, bc\rangle$, contains just a single genotype. The phenotype $\langle ad, bc\rangle$ corresponds to the phenotype $\langle wx\text{-}yz\rangle$ in Φ.

THEOREM 4

Let Φ be a multilocus system, and Trans(Φ) the 1-locus model of Φ. Then Φ has a factor-union representation if and only if Trans(Φ) does.

Proof. Suppose that Φ has a factor-union representation. Let f be the function that assigns sets of factors from a set S to the alleles in Φ. The factor-union representation of Trans(Φ) will use the same set S and a function g similar to f. For each allele $A(\alpha)$ in Trans(Φ), define

$$g(A(\alpha)) = f(\alpha).$$

Since all unions in Trans(Φ) are derived from unions in Φ, it is easy to see that we have constructed a factor-union representation for Trans(Φ).

Deriving a factor-union representation of Φ from one of Trans(Φ) is more involved, since none of the entities of Trans(Φ) correspond directly to the alleles of Φ.

Some additional concepts are needed to conclude the proof. If X is a set of alleles of Φ, let Hom(X) be the set of homozygotes that contain all the elements in X. If X has more than one allele from the same locus, Hom(X) will be empty. Now define

$$F(X) = \prod_{\alpha \text{ in Hom}(X)} g(A(\alpha)).$$

I claim that if X, Y, and Z are sets of alleles of Φ such that $Z = X + Y$ and Hom(Z) is nonempty,

$$F(Z) = F(X) + F(Y) \qquad (*)$$

Since Hom(Z) is a subset of Hom(X) and Hom(Y), both of them are also nonempty. Pick arbitrary χ in X and υ in Y. There is a ζ in Hom(Z) and another homozygote λ such that $\zeta + \lambda$ are the same genotype as $\chi + \upsilon$. For brevity, use μ to denote this genotype. Because g is a factor-union representation,

$$g(A(\zeta)) \leqslant g(A(\mu)) = g(A(\chi)) + g(A(\upsilon))$$

Furthermore,

$$F(Z) \leqslant g(A(\zeta)).$$

Thus

$$F(Z) \leqslant g(A(\chi)) + g(A(\upsilon))$$

for all χ in X and υ in Y. By fixing χ, letting υ vary through all of Y, and taking the intersection of all terms, we get

$$F(Z) \leqslant g(A(\chi)) + F(Y).$$

Now by letting χ vary through all of X and taking the intersection of all

terms produced, we get

$$F(Z) \leqslant F(X) + F(Y)$$

Since $F(Z)$ is $\geqslant F(X)$ and $F(Y)$, we have $F(Z) \geqslant F(X) + F(Y)$. Consequently, (*) follows.

For each allele a in Φ define

$$f(a) = F(\{a\}).$$

Let Alleles(α) be the set of all alleles in the homozygote α. By induction and (*) it follows that

$$F(\text{Alleles}(\alpha)) = \sum_{a \text{ in } \alpha} f(a).$$

But

$$\text{Hom}(\text{Alleles}(\alpha)) = \{\alpha\},$$

so

$$F(\text{Alleles}(\alpha)) = g(A(\alpha)).$$

This means that the "factor sets" assigned to the alleles by f are such that when the appropriate unions are taken, each homozygote receives the same set of factors g assigns to its image in Trans(Φ). Because heterozygotes are unions of homozygotes, it follows that f is a factor-union representation of Φ. ■

I gratefully thank Dr. Kenneth Lange for introducing me to this problem, supplying me with various references, and some very helpful comments on an early draft of this paper. I would also like to thank Dr. Joshua Lederburg for an interesting discussion that led to my cleaning up the section on multilocus systems. At a seminar where I presented these results he indicated that my reduction of 2-locus systems should also hold for multilocus systems. I thought the matter through and produced the proof that this is correct. Finally, I would like to thank the referee for making some suggestions that improved the exposition of the paper.

REFERENCES

1 C. W. Cotterman Factor-union phenotype systems, in *Computer Applications in Genetics* (N. E. Morton, Ed.), Univ. of Hawaii Press 1969, 1-19.

2 G. Markowsky, Combinatorial aspects of lattice theory with applications to the enumeration of free distributive lattices, Ph.D thesis, Harvard Univ., 1973.

3 G. Markowsky, The representation of posets and lattices by sets, *Algebra Universalis* 11:173–192 (1980).

Structural Identification of Large Systems by Reduction to Subsystems: VLDL Triglycerides

J. EISENFELD
Department of Mathematics, University of Texas at Arlington, Arlington, Texas 76010

AND

S. M. GRUNDY
Center for Human Nutrition, University of Texas Health Science Center at Dallas, 5323 Harry Hines Boulevard, Dallas, Texas 75235

Received 30 July 1982; revised 11 March 1983

ABSTRACT

The problem of determining whether or not a particular parameter, or the entire compartmental system, is structural identifiable can be extremely difficult if the number of unknown parameters and the number of compartments is large. However, the problem can be considerably simplified if the system can be decomposed into smaller subsystems in such a way that a parameter is structural identifiable with respect to the large system if and only if it is structural identifiable with respect to the subsystem in which it is contained. The paper offers sufficient conditions under which the desired decomposition can be achieved. The conditions are expressed in terms of the digraph of the system so that they are not difficult to verify. Illustrative examples are provided in terms of application to lipoprotein kinetics.

1. INTRODUCTION

Experiments are performed for identification purposes, i.e. to identify the values of unknown parameters from data. In the event that one or more parameters cannot be identified, the cause could be the result of a variety of problems: insufficient or infrequent sampling, random or nonrandom disturbances, numerical ill conditioning, etc. The input-output configuration may also be the cause of nonidentifiability. In other words, even if the sampling and numerical procedure could be carried out under the most ideal conditions, certain parameters might not be identifiable because these parameters were not uniquely contained in the transfer function, in which case these parameters are said to be not structurally identifiable [1].

©Elsevier Science Publishing Co., Inc., 1983
52 Vanderbilt Ave., New York, NY 10017 0025-5564/83/$03.00

During the past decade a variety of criteria have been established in order to test the structural-identifiability properties of parameters for a general compartmental system. (See [2–9] and references therein.) These tests are analytical, requiring only the model, the set of unknown parameters, and the input-output configuration. Since actual data are not required, these structural-identifiability tests should be performed prior to the planned experiment in order to determine if identifiability is at least possible. In other words, the most appropriate time to consider structural identifiability is during the planning stages of the experiment.

With the advent of the better sampling techniques, sophisticated computer technology, and improved numerical procedures, biological models are becoming increasingly larger and more complex. However, the analytical methods (by themselves) become computationally unfeasible as the structure to be examined increases in complexity. The problems arising from large systems are becoming a major issue in the theory of structural identifiability. One approach to this problem is computerization of the existing methods [2,3]. Another approach is to decompose the large system into smaller systems [4,5]. This paper uses the latter approach. Of course, there are a multitude of ways of decomposing a system into subsystems, and a variety of techniques can be employed to take advantage of the decomposition.

The basic idea, to be presented here, is to represent the system as two systems in series (Figure 1) or as two subsystems in parallel (Figure 2). In either case the transfer function Φ is a combination (a product or a sum) of the transfer functions $\Phi^{(i)}$ $(i=1,2)$ of the constituent subsystems. So the problem is to obtain necessary and sufficient conditions for structurally identifying the constituent transfer functions. Of course the constituent subsystems may be further decomposed.

The main purpose of this paper is to formulate criteria for decomposing a large system into subsystems in such ways as to be able to determine the identifiability properties of parameters, with respect to the large system, by their identifiability properties, with respect to the subsystem in which they are contained. These criteria will be illustrated by applying them to the

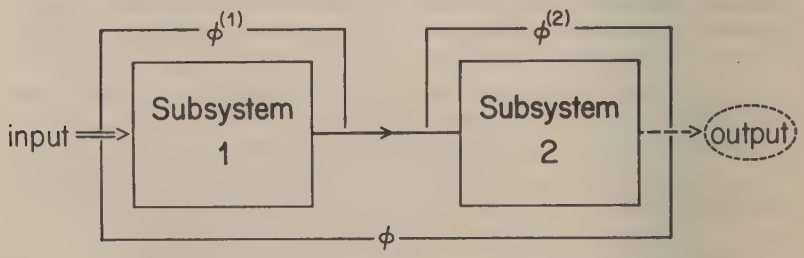

FIG. 1. Subsystems in series: $\Phi = \Phi^{(2)}\Phi^{(1)}$.

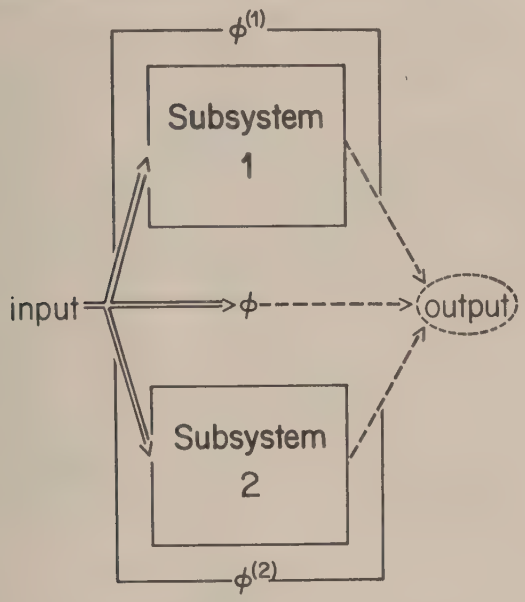

FIG. 2. Subsystems in parallel: $\Phi = \Phi^{(1)} + \Phi^{(2)}$.

model which is used to describe the kinetics of very low-density lipoprotein (VLDL) triglycerides.

Being primarily concerned with applications to tracer and drug studies, we use the nomenclature from compartmental analysis; however, the results apply in generality to the class of linear time-invariant systems.

2. PROBLEM STATEMENT, DEFINITIONS, AND PRELIMINARY RESULTS

A. CONCEPTS FROM SYSTEMS THEORY AND COMPARTMENTAL ANALYSIS

We consider a first-order perturbation experiment with n_B test inputs u_r, and n_C measurement outputs y_m, on an n-compartmental system in a constant steady state. The set of equations representing the dynamics of the deviation x_i of the generic ith compartment from its steady-state value and of the relations between such deviations and measurements is linear and time-invariant [10]:

$$\dot{x}_i = -\sum_{j=0}^{n} a_{ji}x_i + \sum_{j=1}^{n} a_{ij}x_j + \sum_{r=1}^{n_B} b_{ir}u_r \qquad (i=1,2,\ldots,n),$$

$$y_m = \sum_{i=1}^{n} c_{mi}x_i \qquad (m=1,2,\ldots,n_C),$$

or, in the usual system-theory matrix form,

$$\dot{\mathbf{x}} = A\mathbf{x} + B\mathbf{u}, \qquad \mathbf{x}(0) = \mathbf{0}, \qquad \mathbf{y} = C\mathbf{x}, \tag{2.1}$$

where $\mathbf{x}$ is the *state vector* formed by the n variables x_i, $\mathbf{u}$ is the *input vector* formed by the n_B inputs [including instantaneous (delta function) inputs], $\mathbf{y}$ is the *output vector* formed by the n_C experimental measurements, B is the $n \times n_B$ *input matrix* of elements b_{ir}, C is the $n_C \times n$ *output matrix* of elements c_{mi}, and A is the $n \times n$ *compartmental matrix* of *fractional transfer coefficients*, a_{ij} $(i \neq j)$ and $a_{jj} = -\sum_{j=0, j \neq i} a_{ij}$ (a_{0j} denotes the fractional transfer from j to the environment).

The dynamical system (2.1) is characterized by its *transfer function* (*matrix*),

$$\Phi(s) = C(sI - A)^{-1}B, \tag{2.2}$$

where s is the *Laplace transform variable* and I generically denotes an identity matrix. An identity which is useful in connection with the transfer function is [11]

$$(sI - A)^{-1} = \frac{R(s)}{\Delta(s)}, \tag{2.3}$$

where

$$\Delta(s) = \det(sI - A)^{-1} = s^n + \delta_{n-1}s^{n-1} + \delta_{n-2}s^{n-2} + \cdots + \delta_0. \tag{2.4}$$

is the *characteristic polynomial*, and

$$R(s) = \left(s^{n-1} + \delta_1 s^{n-2} + \cdots + \delta_{n-1}\right)I + \left(s^{n-2} + \delta_1 s^{n-3} + \cdots + \delta_{n-1}\right)A + \cdots + (s + \delta_1)A^{n-2} + A^{n-1}. \tag{2.5}$$

Premultiplying and postmultiplying in (2.5) by C and B, respectively, we obtain

$$\Phi(s) = \frac{P(s)}{\Delta(s)}, \tag{2.6}$$

where

$$P(s) = CR(s)B. \tag{2.7}$$

Each entry in the transfer function is a ratio of polynomials,

$$\phi_{ij}(s) = \frac{P_{ij}(s)}{\Delta(s)}. \tag{2.8}$$

If $\Delta(s)$ has a monomial factor which is a common factor of all $P_{ij}(s)$, $i=1,2,\ldots,n_C$, $j=1,2,\ldots,n_B$, then $\Phi(s)$ is *reducible*; otherwise it is *irreducible*. The system (2.1) is reducible or irreducible according as its transfer function is reducible or irreducible.

Example 2.1. The system in Figure 5 (Section 5) arises from decomposing the larger system, Figure 3. The transfer function corresponding to Figure 5 (see Example 4.1) is

$$\Phi(s;\theta)=\frac{\theta_4\theta_6\theta_7\theta_8\theta_9\theta_{10}}{(s+\theta_6)(s+\theta_7)(s+\theta_8)(s+\theta_9)(s+\theta_{10})}+\frac{\theta_5\theta_{11}}{s+\theta_{11}}. \quad (2.9)$$

[$\theta=(\theta_4,\theta_5,\theta_6,\theta_7,\theta_8,\theta_9,\theta_{10},\theta_{11})$ is the parametrization vector.] The characteristic polynomial, $\Delta(s;\theta)$, is the product of the six monomials appearing in (2.9). Let $q(s;\theta)$ be the least common multiple of these monomials. Then $q=\Delta$ if and only if

$$\theta_{11}\neq\theta_k, \qquad k=6,7,8,9,10. \quad (2.10)$$

Thus the condition (2.10) is necessary and sufficient for irreducibility.

Remark 2.1. The irreducibility conditions (2.10) are satisfied at all points in the eight-dimensional parameter space except on the lower-dimensional hyperplanes, $\theta_{11}=\theta_k$, $k=6,7,8,9,10$. Thus $\Phi(s;\theta)$ is irreducible *almost everywhere*. This fact is deduced more easily from input-output connectability and related criteria [7, 8]. We shall say more about the "almost everywhere" property in Remark 3.1.

The system (2.1) is *CC* (*completely controllable*) if its *controllability matrix* $\mathsf{U}=[B,AB,\ldots,A^{n-1}B]$ is of full rank (i.e., $\mathsf{U}\mathsf{U}'$ is nonsingular); it is *CO* (*completely observable*) if its *observability matrix* $V=[C',A'C',\ldots,(A')^{n-1}C']$ ("$'$" means transpose) has full rank. We will require the following well-known result [12; 13, 6]).

LEMMA 2.1

A system is irreducible if and only if it is both CC and CO.

Example 2.2. The system in Figure 4 arises from decomposing the larger system, Figure 3. In this case

$$C,A,B=(1,0),\begin{pmatrix}-(\theta_2+\theta_3+\theta_4+\theta_5) & \theta_1\\ \theta_2 & -\theta_1\end{pmatrix},\begin{pmatrix}1\\0\end{pmatrix},$$

which gives $\det\mathsf{U}=\theta_2$, $\det V=\theta_1$. Since θ_1,θ_2 are nonzero parameters, this system is irreducible.

The matrices which appear in $P(s)$ [Equations (2.7), (2.5)] are the $n_C \times n_B$ matrices

$$M_k = CA^kB, \qquad k = 0,1,\ldots. \tag{2.11}$$

They are called the *Markov matrices*, or the *Markov parameters* in the case $n_C = n_B = 1$. The following result is well known [6].

LEMMA 2.2

All the Markov matrices (2.11) *are uniquely determined from the transfer function. Conversely, the Markov matrices* M_k, $k = 0,1,\ldots,2n-1$ (*the first* $2n$ *suffice*), *uniquely determine the transfer function.*

B. STRUCTURAL IDENTIFICATIONS

We consider a planned experiment. We know, *a priori*, that (for a nonresonant system) the transfer function will have the form [6]

$$\Phi(s) = \sum_{k=1}^{r} \frac{D_k}{s + \pi_k}, \tag{2.12}$$

where the D_k and the π_k are estimable from input-output data, without regard to system parameters. We call the function defined by Equation (2.12) the *measurable transfer function*. On the other hand, a formal expression for the transfer function is obtained from Equation (2.2); however, the matrices which appear in this formula are given in terms of a parameter vector $\boldsymbol{\theta}$, and so the *parametrized transfer function* is

$$\Phi(s;\boldsymbol{\theta}) = C(\boldsymbol{\theta})[sI - A(\boldsymbol{\theta})]^{-1}B(\boldsymbol{\theta}). \tag{2.13}$$

The structural-identifiability problem consists of establishing whether or not $\boldsymbol{\theta}$, or a single parameter θ, can be determined from the relationship

$$\Phi(s;\boldsymbol{\theta}) = \Phi(s). \tag{2.14}$$

By a *single parameter*, we mean a constituent θ_i of θ or a lumped combination of the θ_i. The dimension of $\boldsymbol{\theta}$ is assumed minimal; there are no relations between the θ_i.

There are several useful definitions of structural identifiability; each carries a particular point of view [9]. Some of these are given below.

i. Unique identifiability (also called global identifiability). The single parameter θ is said to be *uniquely identifiable* (from the transfer function) if there exists a unique solution for θ satisfying Equation (2.14). If all components θ_i of the parameter vector $\boldsymbol{\theta}$ are uniquely identifiable, then the model is said to be *parameter-identifiable.*

Example 2.3. Equation (2.13), applied to the system in Example 2.2, yields

$$\Phi(s;\theta)=\frac{s+\theta_1}{s^2+(\theta_1+\theta_2+\sigma)s+\theta_1\sigma}, \tag{2.15}$$

where $\sigma=\theta_3+\theta_4+\theta_5$. Since the 2-compartmental system is irreducible (Example 2.2), the measurable transfer function has the form $\Phi(s)=D_1/(s+\pi_1)+D_2/(s+\pi_2)$ (the π_i are distinct: see [14]). Equation (2.14) yields $\theta_1=D_1\pi_2+D_2\pi_1$, $\theta_1+\theta_2+\sigma=\pi_1+\pi_2$, $\theta_1\sigma=\pi_1\pi_2$. Consequently, θ_1, θ_2, and σ are uniquely identifiable.

*ii. Identifiability (*also called *local identifiability).* If there does not exist a unique solution for a particular parameter but there are at most finitely many solutions, as in Example 2.4 (below), then it may still be possible to ascertain the appropriate solution, say, from *a priori* knowledge of the parameter's range. Therefore, the following definition is useful. The single parameter θ is said to be *identifiable* if there exist at most finitely many solutions for θ satisfying Equation (2.14). If all θ_i are identifiable, the model is said to be *system-identifiable.*

Example 2.4. We return to Example 2.1. Assuming irreducibility (it will be shown later that this condition is not necessary) the measurable transfer function has the form $\Phi(s)=E/(s+\pi_3)(s+\pi_4)(s+\pi_5)(s+\pi_6)(s+\pi_7)+D/(s+\pi_8)$. Equating this expression to (2.9) yields $\theta_{11}=\pi_8$ and the equality between sets $\{\theta_6,\theta_7,\theta_8,\theta_9,\theta_{10}\}=\{\pi_3,\pi_4,\pi_5,\pi_6,\pi_7\}$. Each θ_k, except θ_{11}, has five possible solutions; thus θ_k is identifiable but is *not* uniquely identifiable, while θ_{11} is uniquely identifiable. Finally, θ_4 is identifiable and θ_5 is uniquely identifiable from the relationships $\theta_4=E/(\theta_6\theta_7\theta_8\theta_9\theta_{10})$, $\theta_5=D/\theta_{11}$ (the θ_i are nonzero parameters).

It is interesting that this system is identifiable whether or not the irreducibility condition (2.10) is satisfied (see Example 4.2). The example is among many which show that irreducibility is not necessary for identifiability [6].

iii. Identifiability from a constituent function. Suppose that the (measurable and parametrized) form of the transfer function is expressed in terms of two other functions, e.g., $\Phi=\Phi^{(2)}\Phi^{(1)}$ or $\Phi=\Phi^{(1)}+\Phi^{(2)}$, as depicted in Figures 1 and 2. The relationship is denoted by

$$\Phi(s)=F(f(s),g(s)),\qquad \Phi(s;\theta)=F(f(s;\theta),g(s;\theta)). \tag{2.17}$$

A constituent function f of the transfer function is said to be identifiable (uniquely identifiable) if there exist at most finitely many solutions (a unique solution) for f from the relationship $\Phi(s;\theta)=\Phi(s)$.

The single parameter θ is said to be identifiable (uniquely identifiable) from the constituent function f if there are at most finitely many solutions (a unique solution) for θ from the relationship $f(s;\theta) = f(s)$.

The following results should be apparent from the above definitions.

LEMMA 2.3

Suppose that a constituent function f is identifiable (uniquely identifiable) and that a single parameter θ is identifiable (uniquely identifiable) from f. Then θ is identifiable (uniquely identifiable).

LEMMA 2.4

The characteristic polynomial Δ and the polynomial P (which constitute the transfer function as given in Equation (2.6)) are both uniquely identifiable constituent functions, provided that the system is irreducible.

3. SYSTEMS IN SERIES

In this section we consider a system whose flow diagram is depicted in Figure 1. To be precise we need to distinguish compartments in terms of the input-output relations. Consider the dynamical equations (2.1). A compartment k is said to be an *input compartment* if $b_{kj} \neq 0$ for at least one index j; it is said to be an *output compartment* if $c_{ik} \neq 0$ for at least one index i. We also distinguish the compartments in terms of the direction of flow within the system. A compartmental system is said to admit a *T-R* (*transmitter-receiver*) *decomposition* if the compartments can be separated into disjoint sets T (the *transmitters*) and R (the *receiver*) such that

$$a_{tr} = 0, \qquad t \in T, \quad r \in R. \tag{3.1}$$

The system (2.1) is said to be in *series* if it admits a *T-R* decomposition such that no transmitter compartment is an output compartment, i.e.,

$$c_{it} = 0, \qquad t \in T, \quad i = 1,2,\ldots,n_C, \tag{3.2}$$

and no receiver compartment is an input compartment, i.e.,

$$b_{rj} = 0, \qquad r \in R, \quad j = 1,2,\ldots,n_B. \tag{3.3}$$

In other words, a system is in series if and only if the compartments can be numbered in such a way that the matrices A, B, and C can be partitioned:

$$A = \begin{bmatrix} A^{(1)} & 0 \\ E & A^{(2)} \end{bmatrix}, \qquad B = \begin{bmatrix} B^{(1)} \\ 0 \end{bmatrix}, \qquad C = [0, C^{(2)}]. \tag{3.4}$$

Here "0" generically denotes a matrix of zero entries, $A^{(1)}$ ($A^{(2)}$) is the $n^{(1)} \times n^{(1)}$ ($n^{(2)} \times n^{(2)}$) compartmental matrix whose elements are the fractional transfer coefficients from transmitters to transmitters (receivers to receivers), E is the $n^{(2)} \times n^{(1)}$ matrix whose elements are the fractional transfer coefficients from transmitters to receivers, $B^{(1)}$ is an $n^{(1)} \times n_B$ (input) matrix, and $C^{(2)}$ is an $n_C \times n^{(2)}$ (output) matrix.

LEMMA 3.1

If a system is in series, then the transfer function is a product of constituent transfer functions, i.e., it is expressible in the form

$$\Phi(s) = \Phi^{(2)}(s)\Phi^{(1)}(s), \tag{3.5}$$

where

$$\Phi^{(k)}(s) = C^{(k)}(sI - A^{(k)})^{-1}B^{(k)}, \qquad k = 1,2. \tag{3.6}$$

Proof. Substituting (3.4) into (2.2) yields

$$\Phi(s) = C^{(2)}(sI - A^{(2)})^{-1}E(sI - A^{(1)})^{-1}B^{(1)}. \tag{3.7}$$

The representation (3.5)–(3.6) results from factoring E,

$$E = B^{(2)}C^{(1)}. \tag{3.8}$$

(Such a factoring always exists, say by taking one of the factors to be the identity matrix.)

From the point of view of structural identification, the representation (3.5)–(3.6) is not of interest unless the constituent transfer functions are identifiable. For this reason we consider the case where the dimensions of $B^{(2)}$ and $C^{(1)}$ are $n^{(2)} \times 1$ and $1 \times n^{(1)}$, respectively, in which case

$$C^{(1)} = \left[c_1^{(1)}, c_2^{(1)}, \ldots, c_{n^{(1)}}^{(1)}\right], \qquad B^{(2)} = \left[b_1^{(2)}, b_2^{(2)}, \ldots, b_{n^{(2)}}^{(2)}\right]', \tag{3.9}$$

$$\Phi^{(1)} = \left[\phi_1^{(1)}, \phi_2^{(1)}, \ldots, \phi_{n_B}^{(1)}\right], \qquad \Phi^{(2)} = \left[\phi_1^{(2)}, \phi_2^{(2)}, \ldots, \phi_{n_C}^{(2)}\right]', \tag{3.10}$$

$$\phi_{ij} = \phi_i^{(2)}\phi_j^{(1)}, \qquad i = 1,2,\ldots,n_C, \quad j = 1,2,\ldots,n_B. \tag{3.11}$$

Now the question is, does ϕ_{ij} identify each of its factors, $\phi_i^{(2)}$ and $\phi_j^{(1)}$? The question can be considered in parts by writing Equation (2.8) in the form

$$\phi_{ij}(s) = \frac{\gamma_{ij}e_{ij}(s)}{\Delta(s)}, \tag{3.12}$$

where $e_{ij}(s)$ [and also $\Delta(s)$] is a *monic polynomial* (i.e., the leading coefficient is unity) and $\gamma_{ij} \neq 0$. Similarly,

$$\phi_\nu^{(k)}(s) = \frac{\gamma_\nu^{(k)} e_\nu^{(k)}(s)}{\Delta^{(k)}(s)}, \tag{3.13}$$

where $e_\nu^{(k)}(s)$ is monic and $\gamma_2^{(k)} \neq 0$, $k = 1,2$. Substituting (3.12) and (3.13) into (3.11) yields the identities

$$\gamma_{ij} = \gamma_1^{(2)} \gamma_j^{(1)}, \tag{3.14}$$

$$e_{ij}(s) = e_i^{(2)}(s) e_j^{(1)}(s), \tag{3.15}$$

$$\Delta(s) = \Delta^{(2)}(s) \Delta^{(1)}(s). \tag{3.16}$$

Consider the problem of solving for the factors $\gamma_i^{(2)}$ and $\gamma_j^{(1)}$ in terms of the γ_{ij} from the relations (3.14). In general, this problem is not solvable. However, if just one of the $n_B + n_C$ constants $\gamma_i^{(2)}$ and $\gamma_j^{(1)}$ is known, then the remaining constants are obtainable. In fact, if $\gamma_\mu^{(2)}$ is known, then so are $\gamma_j^{(1)} = \gamma_{\mu j} / \gamma_\mu^{(2)}$, $j = 1,2,\ldots,n_B$. Now fixing j, say $j = 1$, the remaining $\gamma_i^{(2)}$ are obtained from $\gamma_i^{(2)} = \gamma_{i1} / \gamma_1^{(1)}$. A similar arguments applies if one of the $\gamma_j^{(1)}$ is known. Now suppose that one of the entries of the matrix $C^{(2)}B^{(2)}$ is nonzero; say it is the μth entry, τ_μ. It is clear from Equations (2.5)–(2.7) that τ_μ is the leading coefficient of $\phi_\mu^{(2)}(s)$, i.e. $\tau_\mu = \gamma_\mu^{(2)}$. Thus if one of the entries of $C^{(2)}B^{(2)}$ is nonzero and it is known, then all of the $n_B + n_C$ constants $\gamma_i^{(2)}$ and $\gamma_j^{(1)}$ are obtainable from the relations (3.14). This is also true if one of the entries of $C^{(1)}B^{(1)}$ is nonzero and known. The above observations are stated formally in Lemma 3.2.

LEMMA 3.2

The $n_B + n_C$ constants $\gamma_i^{(2)}$ and $\gamma_j^{(1)}$ are uniquely identifiable if either (i) *one of the entries of $C^{(1)}B^{(1)}$ is nonzero and it is known* a priori *or* (ii) *one of the entries of $C^{(2)}B^{(2)}$ is nonzero and it is known* a priori.

LEMMA 3.3

Suppose that the system is irreducible. Then the $n_B + n_C + 2$ polynomials $\Delta^{(k)}(s)$, $e_\nu^{(k)}(s)$ are identifiable.

Proof. Since the system is irreducible, $\Delta(s)$ and $P_{ij}(s)$, $1 \leqslant i \leqslant n_C$, $1 \leqslant j \leqslant n_B$, are uniquely identifiable (Lemma 2.4). Let $\pi_j^{(1)}$ be a root of $\Delta^{(1)}(s)$. Then, from (3.16), $\pi_j^{(1)}$ is a root of $\Delta(s)$; thus there are finitely many solutions for $\pi_j^{(1)}$. This being true for each root of $\Delta^{(1)}(s)$, $\Delta^{(1)}(s)$ is identifiable.

Similar arguments show that $\Delta^{(2)}(s)$ and the polynomials $e_\nu^{(k)}(s)$ are identifiable [recall that $P_{ij}(s)$ and $e_{ij}(s)$ have identical roots].

LEMMA 3.4

Suppose that there is a compartment f, called the first receiver, such that $f \in R$ and $a_{rt} = 0$ whenever $r \in R - f$ and $t \in T$. It is assumed, without loss of generality, that f is numbered first among the receivers. Then the matrices in (3.9) *are*

$$C^{(1)} = \left[a_{f1}, a_{f2}, \ldots, a_{fn^{(1)}}\right], \qquad B^{(2)} = [1,0,\ldots,0]'. \tag{3.17}$$

Proof. The hypothesis states that all rows of E are zero except for the first row, which is $[a_{f1}, a_{f2}, \ldots, a_{fn^{(1)}}]$. Clearly, $E = B^{(2)}C^{(1)}$, where $C^{(1)}$ and $B^{(2)}$ are defined in (3.17).

Example 3.1. We consider the system in Figure 5(b) (Section 5). The system has a T-R decomposition where $T = \{10\}$ and $R = \{11,12,13,14\}$. The matrices in Equation (3.4) are

$$B^{(1)} = [\theta_4], \qquad C^{(2)} = [0,0,0,\theta_{10}]', \qquad A^{(1)} = [-\theta_6],$$

$$E = \begin{bmatrix} \theta_6 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \qquad A^{(2)} = \begin{bmatrix} -\theta_7 & 0 & 0 & 0 \\ \theta_7 & -\theta_8 & 0 & 0 \\ 0 & \theta_8 & -\theta_9 & 0 \\ 0 & 0 & \theta_9 & -\theta_{10} \end{bmatrix}.$$

The system has a first receiver $f = 11$, and so $E = B^{(2)}C^{(1)}$, where $C^{(1)} = [\theta_6]$ and $B^{(2)} = [1,0,0,0]'$. By Lemma 3.4, $\Phi = \Phi^{(2)}\Phi^{(1)}$, where $\Phi^{(1)}(s) = \theta_4\theta_6/(s + \theta_6)$. Now the receiver constituent is also a system in series with a first receiver. In fact, Lemma 3.4 can be applied in succession to obtain

$$\Phi(s) = \frac{\theta_4\theta_6}{s+\theta_6}\,\frac{\theta_7}{s+\theta_7}\,\frac{\theta_8}{s+\theta_8}\,\frac{\theta_9}{s+\theta_9}\,\frac{\theta_{10}}{s+\theta_{10}}. \tag{3.18}$$

where each fraction is a constituent transfer function. It is apparent from the form of (3.18) that the poles θ_k, $k = 6,7,8,9,10$, are identifiable and so is the constant $\theta_4\alpha$, where α is the product of the poles. Since α is identifiable, so is θ_4. Thus all six parameters are identifiable.

LEMMA 3.5

Suppose that there is a compartment l, called the last transmitter, such that $l \in T$ and $a_{rt} = 0$ whenever $r \in R$ and $t \in T - l$. It is assumed, without loss of generality, that l is numbered last among the transmitters. Then the matrices in (3.9) *are*

$$C^{(1)} = [0,\ldots,0,1], \qquad B^{(2)} = [a_{1l}, a_{2l}, \ldots, a_{n^{(2)}l}]'. \tag{3.19}$$

Proof. The hyperthesis states that all columns of E are zero except for the last column, which is $[a_{1l}, a_{2l}, \ldots, a_{n^{(2)}l}]'$. Clearly, $E = B^{(2)}C^{(1)}$, where $C^{(1)}$ and $B^{(2)}$ are defined in (3.19).

Example 3.2. We consider the system in Figure 7 (Section 5). It is a system in series with $l = 4$ as the last transmitter. The system decomposes into a transmitter subsystem (Figure 4) and a receiver subsystem (Figure 5).

PROPOSITION 3.1

Suppose that a system in series has the following properties:

(1) *There is either* (a) *a first receiver or* (b) *a last transmitter.*
(2) *One of the conditions* (i) *or* (ii) *in Lemma* 3.2 *holds.*
(3) *The system is irreducible.*

The the transfer function $\Phi = \Phi^{(2)}\Phi^{(1)}$, *where the constituent transfer functions are given in Equations* (3.6), (3.4), *and* (3.17) *for the first-receiver case, or in Equations* (3.6), (3.4), *and* (3.19) *for the last-transmitter case. Moreover, the constituent transfer functions are identifiable.*

Proof. The decomposition $\Phi = \Phi^{(2)}\Phi^{(1)}$ follows from Lemma 3.1. To obtain the first-receiver form (3.17) we apply Lemma 3.4, and the last-transmitter form (3.19) follows from Lemma 3.5. The fact that the constituent transfer functions are identifiable is a consequence of Lemmas 3.2 and 3.3.

Example 3.3. We continue from Example 3.2. In the transmitter subsystem, Figure 4, there is a single input and a single output, and both occur in compartment 4. Moreover, $C^{(1)}B^{(1)} = [1,0][1,0]' = 1$, so that the second condition in Proposition 3.1 is satisfied. Next we find the conditions under which Φ is irreducible. We recall that Φ is (always) irreducible (Example 2.2) and Φ_2 is irreducible under the condition (2.10) (Example 2.1). Let the condition (2.10) be satisfied; then Φ is reducible if and only if a zero of one of the constituent transfer functions coincides with a pole of the other. Now $\Phi^{(1)}$ is given in Equation (2.15) and $\Phi^{(2)}$ is given in Equation (2.9). There are no cancellations between these functions if and only if

$$\theta_1 \neq \theta_k, \qquad k = 6,7,8,9,10,11, \tag{3.20}$$

$$(\theta_{11} + \pi_i)\theta_4 \prod_{k=6}^{10} \theta_k + \theta_6\theta_{11} \prod_{k=6}^{10} (\theta_k + \pi_i) \neq 0, \qquad i = 1,2, \tag{3.21}$$

where π_1 and π_2 are the roots of

$$\Delta^{(2)}(s) = s^2 + (\theta_1 + \theta_2 + \theta_3 + \theta_4 + \theta_5)s + \theta_1(\theta_3 + \theta_4). \tag{2.22}$$

In conclusion, $\Phi^{(1)}$ and $\Phi^{(2)}$ are identifiable as long as the irreducibility condition (2.10), (3.20), and (3.21) are satisfied.

Remark 3.1. The irreducibility conditions may not be easily derivable, especially in large systems, and moreover, it may be difficult to determine *a priori* whether or not these conditions are satisfied. It is possible to avoid the issue by adopting the point of view that structural identifiability almost everywhere is a sufficient criterion, because in this case the chance of encountering a point at which the system is not structurally identifiable (in an actual experiment) is negligible [7, 8]. From this point of view, Proposition 3.1 may be modified to read as follows.

PROPOSITION 3.2

If hypothesis (3) *of Proposition* 3.1 *is changed to*

(3′) *the system is irreducible almost everywhere,*

then the conclusions of that proposition hold in the "almost everywhere" sense.

Moreover, "irreducible almost everywhere" criteria are less difficult to apply. For instance, in all the examples discussed above, the system satisfies the following conditions:

(1) Every compartment is input-reachable.
(2) Every compartment reaches at least one output.
(3) The system is without traps.

Systems satisfying these three conditions are irreducible almost everywhere [7, 8].

Remark 3.2. Even though one adopts the "almost everywhere" point of view, the irreducibility conditions are still of significance. For instance, some computer programs assume irreducibility (explicitly or implicitly [18]). In such cases, numerical difficulties can arise (in the estimation problem) in proximity to a point where irreducibility fails. The numerical difficulties become apparent when we expand the impulse response of the compartmental system as the exponential sum $D_1 e^{\pi_1 t} + D_2 e^{\pi_2 t} + \cdots + D_n e^{\pi_n t}$. If the order of the system is n, then we expect to have n linearly independent terms, i.e. (i) the π_i are distinct and (ii) the D_i are nonzero. A necessary and sufficient condition for conditions (i) and (ii) to hold is that the system is irreducible. As the parameter vector approaches a point where the irreducible conditions are violated, then either some of the π_i coalesce or some of the D_i tend to zero; either one of these circumstances is apt to cause numerical difficulties in the parameter-estimation problem.

4. SYSTEMS IN PARALLEL

In this section we consider a system whose flow diagram is depicted in Figure 2 and which is described below.

The system (2.1) is said to be *in parallel* if its compartments can be separated into two disjoint subsets $S^{(k)}$, $k=1,2$, such that there is no flow between these subsets, i.e.,

$$a_{ij}=0, \quad i\in S^{(1)}, \quad j\in S^{(2)}. \tag{4.1}$$

In other words, a systems is in parallel if and only if the compartments can be numbered in such a way that the compartmental matrix can be partitioned as

$$A=\begin{bmatrix} A^{(1)} & 0 \\ 0 & A^{(2)} \end{bmatrix}, \tag{4.2}$$

where $A^{(1)}$ and $A^{(2)}$ are submatrices of dimensions $n_1\times n_1$ and $n_2\times n_2$, respectively. We also partition B and C as

$$B=[B^{(1)},B^{(2)}]', \qquad C=[C^{(1)},C^{(2)}], \tag{4.3}$$

where $B^{(1)}$, $B^{(2)}$, $C^{(1)}$, and $C^{(2)}$ have respective dimensions $n_1\times n_B$, $n_2\times n_B$, $n_C\times n_1$, and $n_C\times n_2$. By substituting (4.2)–(4.3) into (2.2) we obtain the following result.

LEMMA 4.1

A system in parallel has the property that its transfer function is expressible as a sum of constituent transfer functions,

$$\Phi(s)=\Phi^{(1)}(s)+\Phi^{(2)}(s), \tag{4.4}$$

where

$$\Phi^{(k)}(s)=C^{(k)}(sI-A^{(k)})^{-1}B^{(k)}, \qquad k=1,2. \tag{4.5}$$

Example 4.1. The system in Figure 5 (Section 5) is in parallel. The constituent subsystems are depicted in Figure 5(a) and (b). The transfer function for the "short" branch is $\Phi^{(2)}(s)=\theta_5\theta_{11}/(s+\theta_{11})$. The other branch was discussed earlier (Example 3.1); its transfer function is given in Equation (3.18). The transfer function is the sum of its constituent transfer functions as given by Equation (2.9).

Certain concepts from the theory of directed graphs (e.g., see [15]) are needed. Let D denote the *digraph* (directed graph) of a compartmental matrix A, i.e., a directed arc is drawn from compartment j to (another) compartment i if and only if $a_{ij}\neq 0$. A *sequence* from (compartment) j to i,

$$\Gamma: j=\nu_0\to\nu_1\to\cdots\to\nu_l=i,$$

exists in D if and only if the product

$$\pi(\Gamma) = a_{\nu_0\nu_1} a_{\nu_1\nu_2} \cdots a_{\nu_{l-1}\nu_l} \neq 0.$$

The integer l is the *length* of the sequence. A *path* is a sequence which does not intersect itself, i.e., $\nu_s \neq \nu_t$ if $s \neq t$.

LEMMA 4.1

Let (A, B, C) be a system (as described in Section 1). Suppose that p is a positive integer such that each path from an input compartment to an output compartment has length $> p$. Then the Markov matrices

$$M_k = CA^kB = 0, \qquad k = 0,1,\ldots,p.$$

Proof. The (i, j) entry of M_k is a sum of products of the form $c_{i\nu_0} a_{\nu_0\nu_1} \cdots a_{\nu_{k-1}\nu_k} b_{\nu_k j}$. For $M_k \neq 0$ there must be an input compartment ν_0 and an output compartment ν_k such that there is a sequence from ν_0 to ν_k of length k. Certainly, there must exist a path from ν_0 to ν_k of length k. Consequently, $k > p$, which is the desired result.

PROPOSITION 4.1

Suppose that a system in parallel has the property that the number of compartments $n^{(2)}$, in one constituent, say $(A^{(2)}, B^{(2)}, C^{(2)})$, is sufficiently small so that each path in the other constituent, $(A^{(1)}, B^{(1)}, C^{(1)})$, from an input compartment to an output compartment has length $> 2n^{(2)} - 1$. Then each constituent transfer function is uniquely identifiable.

Proof. By the hypothesis and Lemma 4.1,

$$M_k^{(1)} = C^{(1)}(A^{(1)})^k B^{(1)} = 0, \qquad k = 0,1,\ldots,2n^{(2)} - 1.$$

Since

$$M_k = CA^kB = [C^{(1)}, C^{(2)}] \begin{bmatrix} (A^{(1)})^k & 0 \\ 0 & (A^{(2)})^k \end{bmatrix} \begin{bmatrix} B^{(1)} \\ B^{(2)} \end{bmatrix}$$

$$= C^{(1)}(A^{(1)})^k B^{(1)} + C^{(2)}(A^{(2)})^k B^{(2)} = M_k^{(1)} + M_k^{(2)},$$

$M_k^{(2)} = M_k$ is uniquely identifiable, $k = 0,1,\ldots,2n^{(2)} - 1$. In view of Lemma 2.2, $\Phi^{(2)}$ is uniquely identifiable and hence $\Phi^{(2)} = \Phi - \Phi^{(1)}$ is also uniquely solvable.

Example 4.2. Let Φ be the transfer function for the system in Figure 5. The "short" branch has one compartment, i.e., $n^{(2)} = 1$, while the length of

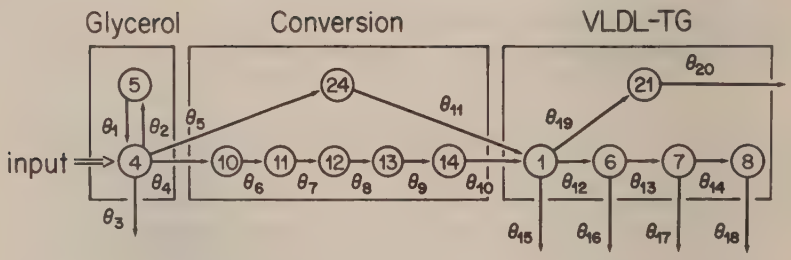

FIG. 3. VLDL-TG kinetics. The tracer is injected into compartment 4 (plasma). The observed VLDL-TG tracer curve is $y = x_1 + x_6 + x_7 + x_8 + x_{21}$.

the path (there is only one) from the input compartment to the output compartment in the larger branch is 4. By Proposition 4.1, each constituent transfer function is uniquely identifiable from Φ. Now $\Phi^{(2)}(s) = \theta_5\theta_{11}/(s + \theta_{11})$ uniquely identifies θ_5 and θ_{11}, while $\theta_4, \theta_6, \theta_7, \theta_8, \theta_9, \theta_{10}$ are identifiable from $\Phi^{(1)}$ (Example 3.1). Notice that all the parameters in Φ are identifiable from Φ, whether or not the irreducibility condition (2.10) is satisfied.

5. VERY LOW-DENSITY LIPOPROTEIN TRIGLYCERIDES

A. MODEL DESCRIPTION

The kinetics of the triglyceride (TG) moiety of plasma very low-density lipoproteins (VLDL) in man, as discussed by Zech et al. [16], Berman [17], and Beltz et al. [18], may be summarized as follows: Radioactively labeled glycerol is injected into the plasma to serve as the glycerol backbone of the TG moiety; the resulting VLDL-TG tracer curve is observed for forty-eight hours. The compartmental model, Figure 3, may be described in parts.

The free glycerol precursor subsystem, Figure 4, describes the kinetics of intravenously injected glycerol in plasma. Labeled glycerol (G^*) entering the

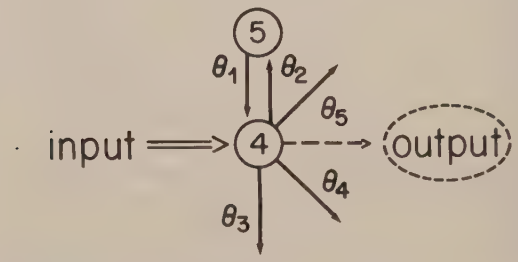

FIG. 4. Free glycerol percursor subsystem. Compartment 4 is observed (indirectly, as a consequence of Lemma 3.5).

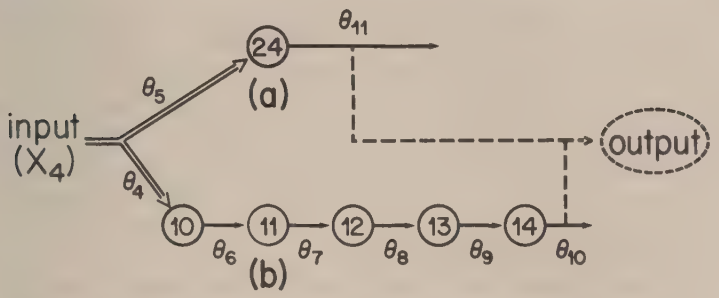

FIG. 5. Conversion subsystem: (a) slow conversion pathway, (b) fast conversion pathway. $y = \theta_{10}x_{14} + \theta_{11}x_{24}$ is observed (indirectly, as a consequence of Lemmas 3.4 and 3.5).

plasma (compartment 4) equilibrates with the glycerol pool in a two-phase decay process involving extravascular (compartment 5) exchange. Malmendier et al. [19] found that about 90% of the glycerol leaving the plasma could be accounted for by conversion to glucose and CO_2, leaving only a small fraction for TG synthesis.

The conversion of glycerol into the TG derivative is an esterification reaction in which three fatty acids attach to a glycerol molecule with loss of H_2O. The incorporation of glycerol into TG follows two pathways as shown in Figure 5. Synthesized TG then reacts with cholesterol and protein to form plasma TG-VLDL.

Plasma VLDL-TG (Figure 6) consists of a delipidation chain (1,6,7,8) and a single compartment (21), which represents a more slowly catabolized subfraction of VLDL particles. The delipidation chain represents a spectrum of decreasing VLDL-TG particles. Of the triglyceride secreted with the largest particles, a portion is removed during the delipidation, which occurs as the particle gets smaller. The remainder of the particle's TG accompanies it to the next step of the catabolic cascade. The smallest VLDL particles

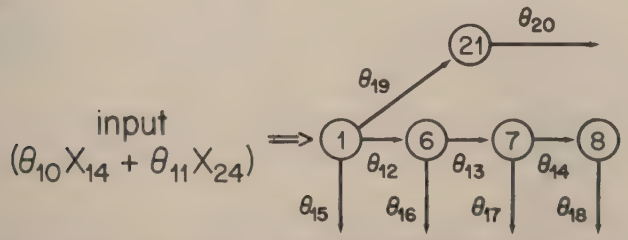

FIG. 6. VLDL-TG subsystem. VLDL-TG inputs into compartment 1, and $y = x_1 + x_6 + x_7 + x_8 + x_{21}$ is observed.

(compartment 8) leaving the delipidation are in the intermediate-density lipoprotein (IDL) range. The portions removed during delipidation convert back to glycerol and fatty acid. The converted glycerol may recycle through the system, but this recycling has been found to have a negligible effect on the overall kinetics [16], and its has been omitted in more recent discussions of the model [18].

B. *STRUCTURAL IDENTIFICATION*

It is interesting that the decomposition of the total system (Figure 3), for the purpose of structural identification, recovers the same subsystems that were used to construct the model in the first place. The first step of the decomposition is achieved by expressing the system in series, with compartment 1 as the first receiver. The resulting receiver subsystem is the VLDL-TG subsystem (Figure 6), and the transmitter system is the synthesis subsystem (Figure 7).

The decomposition of the synthesis subsystem (Figure 5) into constituents, [Figures 4, 5(a), and 5(b)] has already been discussed in Examples 3.2 and 4.1, and its was found (Example 3.3) that the transfer function for the synthesis subsystem, $\Phi^{(S)}$, is a product of the constituent transfer functions given in Equations (2.9) and (2.15). The transfer function for the VLDL-TG subsystem is

$$\Phi^{(V)}(s) = \frac{1}{s+\theta_{12}+\theta_{15}+\theta_{19}} \times \left\{1+\frac{\theta_{12}}{s+\theta_{13}+\theta_{16}}\left[1+\frac{\theta_{13}}{s+\theta_{14}+\theta_{17}}\left(1+\frac{\theta_{14}}{s+\theta_{18}}\right)\right]+\frac{\theta_{19}}{s+\theta_{20}}\right\}. \tag{5.1}$$

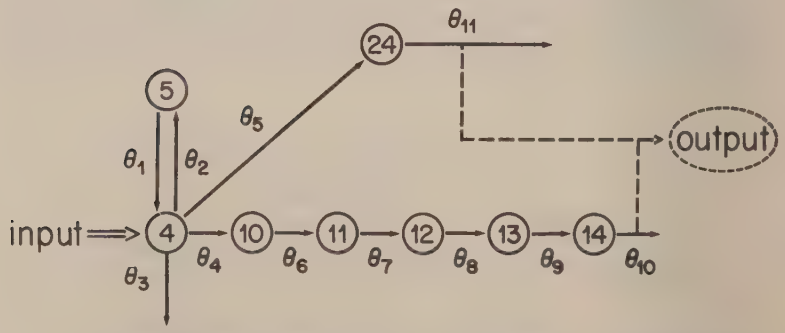

FIG. 7. Synthesis subsystem. The input is into compartment 4, and $y = \theta_{10}x_{14} + \theta_{11}x_{24}$ is observed.

The irreducibility condition for this subsystem is

$$\theta_{20} \notin \{\theta_{13}+\theta_{16}, \theta_{14}+\theta_{17}, \theta_{18}\}, \qquad \theta_{17} \neq \theta_{18}. \tag{5.2}$$

We recall, from Example 3.3., that the synthesis subsystem is irreducible if and only if the conditions (2.10), (3.20), and (3.21) are satisfied. To these we add the conditions which prohibit coincidence between the roots of one constituent transfer function with the zeros of the other; these are

$$\begin{aligned}&(s+\theta_{13}+\theta_{16})(s+\theta_{14}+\theta_{17})(s+\theta_{18})(s+\theta_{20})\\&\quad+\theta_{12}(s+\theta_{14}+\theta_{17})(s+\theta_{18})(s+\theta_{20})\\&\quad+\theta_{12}\theta_{13}(s+\theta_{18})(s+\theta_{20})+\theta_{12}\theta_{13}\theta_{14}(s+\theta_{20})\\&\quad+\theta_{19}(s+\theta_{13}+\theta_{16})(s+\theta_{14}+\theta_{17})(s+\theta_{18})\neq 0\end{aligned}$$

$$\text{for} \quad s=-\theta_k,\ 6\leqslant k\leqslant 11 \ \text{ or } \ s=-\pi_i,\ i=1,2, \tag{5.3}$$

$$\theta_1 \neq \theta_{12}+\theta_{15}+\theta_{17}, \theta_{13}+\theta_{16}, \theta_{14}+\theta_{17}, \theta_{18}, \theta_{20}, \tag{5.4}$$

$$(s+\theta_{11})\theta_4\prod_{k=6}^{10}\theta_k+\theta_6\theta_{11}\prod_{k=1}^{6}(s+\theta_k)\neq 0$$

$$\text{for} \quad s=-\theta_{12}-\theta_{15}-\theta_{19}, -\theta_{13}-\theta_{16}, -\theta_{14}-\theta_{17}, -\theta_{18}, -\theta_{20}. \tag{5.5}$$

Let us assume that all these irreducibility conditions, (2.10), (3.20), (3.22), (5.3), (5.4), and (5.5), are satisfied. Condition (i) in Lemma 3.2 holds because $C^{(V)}B^{(V)}=1$ ("V" refers to the VLDL-TG subsystem). Thus, by Proposition 3.1, the constituent transfer functions $\Phi^{(V)}$ and $\Phi^{(S)}$ are identifiable. It is not difficult to show that the parameters θ_k, $12\leqslant k\leqslant 20$, are identifiable from $\Phi^{(V)}$; thus, these parameters are identifiable (Lemma 2.3).

Since $\Phi^{(S)}$ is identifiable, its constituent transfer functions $\Phi^{(G)}$ and $\Phi^{(C)}$ are identifiable, as shown in Example 3.3 ("G" denotes the free glycerol subsystem, and "C" denotes the conversion subsystem). Now θ_k, $4\leqslant k\leqslant 11$, are identifiable from $\Phi^{(C)}$ (Example 4.2). Moreover, θ_1, θ_2, and $\theta_3+\theta_4$ are identifiable from $\Phi^{(G)}$ (Example 2.3). Since θ_4 is identifiable, so is θ_3. Thus, the system of Figure 3 is identifiable provided that the irreducibility conditions given above are satisfied.

C. *DISCUSSION*

The main objective of structural-identifiability analysis is to ascertain the theoretical limits of the parameter-estimation problem. In practice, the estimation of twenty parameters from a single tracer curve is not feasible, even though the data might be free of noise. Usually, thirteen relationships

between the parameters are stipulated *a priori*, leaving only seven free parameters to be estimated [18].

The irreducibility conditions, (2.10), (3.20), (3.21), (5.3), (5.4), and (5.5), guarantee system identifiability (whether or not *a priori* relationships are imposed). The fact that the irreducibility conditions are satisfied almost everywhere (see Remark 2.1) may also be deduced from the criteria specified in Remark 3.1. Using Proposition 3.2 in place of Proposition 3.1, we obtain system identifiability in the "almost everywhere" sense (without regard to the irreducibility conditions). The irreducibility conditions may also play a role in the parameter-estimation problem (see Remark 3.2).

REFERENCES

1 R. Bellman and K. J. Astrom, On structural identifiability, *Math. Biosci.* 7:329–339 (1970).

2 R. F. Brown and J. P. Norton, Identifiability of large compartmental models, *Large Scale Systems* 3:159–175 (1982).

3 H. Pohjanpalo and B. Wahlstrom, Software for solving identifications and identifiability problems e.g. in compartmental systems, *Math. Comput. Simulation* 24 (1982).

4 M. Milanese and N. Sorrentino, Decomposition methods for the identifiability analysis of large systems, *Internat. J. Control* 28:71–79 (1978).

5 J. Eisenfeld, New techniques for structural identifiability for large linear and nonlinear compartmental systems, *Math. Comput. Simulation* 24:494–501 (1982).

6 J. Eisenfeld, On identifiability of impulse-response in compartmental systems, *Math. Biosci.* 47:15–23 (1979).

7 C. Cobelli and G. Romanin-Jacur, On the structural identifiability of biological compartmental systems in a general input-output configuration, *Math. Biosci.* 30:139–151 (1976).

8 C. Cobelli, A. Lepschy, and G. Romanin-Jacur, Identifiability of compartmental systems and related structural properties, *Math. Biosci.* 44:1–18 (1979).

9 C. Cobelli and J. J. Distefano, III, Parameter and structural identifiability concepts and ambiguities: A critical review and analysis, *Amer. J. Physiol.* 239 (*Regulatory Integrative Comp. Physiol.* 8):R7–R24 (1980).

10 J. A. Jacquez, *Compartmental Analysis in Biology and Medicine*, Elsevier, Amsterdam, 1972.

11 F. R. Gantmacher, *The Theory of Matrices*, Chelsea, New York, 1959; Vol. 2, 1964, pp. 97–98.

12 R. W. Brockett, *Finite Dimensional Linear Systems*, Wiley, New York, 1970.

13 C. T. Chen, *Introduction to Linear Systems Theory*, Holt, Rinehart, and Winston, New York, 1970.

14 J. Z. Hearon, Theorems on linear systems, *Ann. N.Y. Acad. Sci.* 108:36–68 (1963).

15 F. Harary, R. Z. Norman, and D. Cartwright, *Structural Models: An Introduction to the Theory of Directed Graphs*, Wiley, New York, 1965.

16 L. A. Zech, S. M. Grundy, D. Steinberg, and M. Berman, Kinetic model for production and metabolism of very low density lipoprotein triglycerides, *J. Clin. Invest.* 63:1262–1273 (1979).

17 M. Berman, Analysis of kinetic data, *Prog. Biochem. Pharmacol.* 15:67–108 (1979).

18 W. F. Beltz, S. M. Grundy, and T. E. Carew, An efficient program for estimation of kinetic parameters in a model of very low density lipoprotein triglyceride metabolism, in *Lipoprotein Kinetics and Modeling* (M. Berman, S. M. Grundy, and B. Howard, Eds.), Academic, New York, 1982.

19 C. L. Malmendier, C. Delcroix, and M. Berman, Interrelations in the oxidative metabolism of free fatty acids, glucose, and glycerol in normal and hyperlipermic patients, *J. Clin. Invest.* 54:461–476.

Errata

Richard FitzHugh, Statistical Properties of the Asymmetric Random Telegraph Signal, With Applications to Single Channel Analysis, *Math. Biosci.* 64:75–89 (1983)

On page 86, revise equations as follows:

$$h(y_0, y_1, t)\, dy_1 = \text{Prob}[\, y(t_0 + t) \varepsilon (y_1, y_1 + dy_1) | y(t_0) = y_0]$$

$$h(y_0, y_1, t) = \{ q_0 f_0(y_0)[r_{00}(y_0, y_1, t) + r_{01}(y_0, y_1, t)] + q_1 f_1(y_0)[r_{11}(y_0, y_1, t) + r_{10}(y_0, y_1, t)] \} / f(y_0). \tag{23}$$

$$h(y_0, y_1, t) = [q_0 f_0(y_0) \lambda^{u_0} \delta(y_1 - m_0) + q_1 f_1(y_0) \lambda^{u_1} \delta(y_1 - m_1)] / f(y_0) + [q_0 f_0(y_0) u_0 + q_1 f_1(y_0) u_1] / [af(y_0)] + 0(t) \tag{24}$$

$$v_0 v_1 \tau / (va) + 0(t). \tag{25}$$

©Elsevier Science Publishing Co., Inc., 1983
52 Vanderbilt Ave., New York, NY 10017 0025-5564/83/$03.00

BOOK REVIEW

N. Alvey, N. Galwey, and P. Lane, *An Introduction to Genstat*, Academic Press, New York, 1982, 152 pp., $16.00, paperback

---

Eventually, most researchers are confronted with the necessity of analyzing data, perhaps to estimate parameters in a model, or simply to summarize information in a sample. A number of options are available: hand calculations; a special purpose computer program written in ALGOL, FORTRAN, or similar language; or a prewritten general purpose package of statistical computer programs such as SAS, SPSS, IMSL, or Genstat.

Genstat is a flexible command language which simplifies the programming of a wide range of statistical analyses. Commands invoke sophisticated techniques which can generally cope with difficulties such as missing observations not often tolerated by statistical packages. Single commands produce $X-Y$ plots, analyses of designed experiments, multiway tables, cluster analyses, linear regression analyses, and more. Macros written in the Genstat language extend the package to enable more complex analyses such as nonparametric testing, generalized linear model fitting, and estimation of log-dose response.

An Introduction to Genstat assumes no previous knowledge of computers or statistics. The first chapter describes methods for data transfer between the user and Genstat, and commands for simple calculations. Later chapters deal with simple versions of plotting, linear regression, tabulation, and analysis of designed experiments. There are brief sections on programming and debugging, and data storage and retrieval. The last chapter alludes to the potential for more complex analyses. Throughout, the style is clear and systematic. The reader is led through a logical hierarchy of commands. As each new command is introduced, its use is illustrated with an example. Each chapter concludes with problems for which solutions are provided.

In striving for its goal of introducing a statistical package, *An Introduction to Genstat* is severely hampered by the frequent awkwardness and apparent artificiality of Genstat. The reader is overwhelmed by the multitude of details of data structures and language syntax; one must, for instance, differentiate between VARIATES or vector variables and SCALARS. A summary of newly introduced commands is provided at the end of each chapter, but this summary is not complete. The reader must scramble through pages of text in order to put together a program. An appendix of Genstat commands grouped by function would be most useful. Unfortunately, *An Introduction to Genstat* does little to dispel the mystery surrounding Genstat's rules. Why should '$' rather than ';' be used to separate items in a list? However, the book does explain the sometimes obscure origins of some of the Genstat options. Who would guess that $NUN = Q$ tells Genstat not to expect an END OF DATA when reading input?

©Elsevier Science Publishing Co., Inc., 1983
52 Vanderbilt Ave., New York, NY 10017 0025-5564/83/$03.00

As promised, *An Introduction to Genstat* does not assume previous statistical knowledge, and therefore explains the basic concepts when necessary. This approach is not appropriate to Genstat, which has been described as a "statistician's statistician" and can be a dangerous tool in the hands of the statistically naive user. Perhaps the explanations are brief enough to be harmless.

The pace of the book is slow to the point that some readers will feel that it takes too long to begin programming a realistic analysis. However, this criticism is a reflection of the complexity of Genstat rather than a fault of the book. Completing an analysis on Genstat may be frustratingly difficult for a novice to statistical packages, but *An Introduction to Genstat* is a commendable attempt to ease the pain.

In a broader context, one may question the general usefulness of Genstat. Genstat emphasises an empirical analysis of large sets of data. For a mathematical modeler using general functions to represent phenomenological or mechanistic theories, most of the directives will be superfluous. However, the directive OPTIMIZE maximizes likelihood functions under one of three error structures (normal, multinomial, or poisson) and minimizes general functions. Directives for $X-Y$ and contour plotting may also be useful. However, computer installations which might consider implementing Genstat will typically already handle these tasks well using special purpose packages. An additional consideration is that Genstat is generally invoked in batch mode. Analyses lack the spontaneous and exploratory nature of analyses performed interactively, and produce volumes of output which are often complicated to interpret.

Thus, the usefulness of Genstat is very much open to debate. To utilize the power of Genstat, the newcomer must invest considerable time learning the language. Otherwise, one is restricted to simple commands which could easily be performed on a user-friendly but more restricted system such as interactive MINITAB (Pennsylvania State University). However, for those who need the sophisticated statistical routines of Genstat, *An Introduction to Genstat* may prove to be an indispensable tool.

CATHERINE A. MACKEN
Lincoln College
Canterbury, New Zealand

Books Received

Matrix Algebra Useful for Statistics. Shayle R. Searle. In the Wiley series: Probability and Mathematical Statistics. John Wiley and Sons, New York, 1982, xxii + 438 pp., $29.95.

Linear Models: *An Introduction*. Irwin Guttman. In the Wiley series: Probability and Mathematical Statistics. John Wiley and Sons, New York, x + 356 pp., $36.95.

Applicable Mathematics of Non-Physical Phenomena. F. Oliveira-Pinto and B. W. Conolly. John Wiley and Sons, New York, 269 pp., $59.95.

An Introduction to Genstat. N. Alvey, N. Galwey, and P. Lane. Academic Press, New York, 1982, viii + 152 pp., $16.00, paperback.

Mathematical Models in Biological Oceanography. T. Platt, K. H. Mann, and R. E. Ulanowicz, eds. Unesco Press, Paris, 1981, 156 pp., hardback.

Quantitative Approaches to Metabolism: *The Role of Tracers and Models in Clinical Medicine*. D. G. Cramp, ed. John Wiley and Sons, 1982, 390 pp., $59.95, hardback.

Biomathematics: 1*st and* 2*nd Portonovo Conferences on Endocrine Pharmacology*, *Compartmental Models and Control Systems*. Carlo de Martinis and Luigi Rossine, eds. Piccin Medical Books, Padova, Italy, 1981, 319 pp., $25.00, paperback.

Multivariate Data Analysis in Industrial Practice. Paul J. Lewi. John Wiley and Sons, New York, 1982, xi + 244 pp., $31.95.

A Second Course in Stochastic Processes. Samuel Karlin and Howard M. Taylor. Academic Press Inc., New York, 1981, xviii + 542 pp., $35.00.

Introduction to Probability Models, 2nd Edition. Sheldon M. Ross. Academic Press, New York, 1981, xi + 376 pp., $19.25.

An Introduction to Applied Optimal Control. Greg Knowles. Academic Press, New York, 1981, $19.50.

An Introduction to Microcirculation. Mary P. Wiedeman, Ronald F. Tuma, and Harvey N. Mayrovitz. Academic Press, New York, 1981, xi + 226 pp., $30.00.

Practical Optimization. Philip E. Gill, Walter Murray, and Margaret H. Wright. Academic Press, New York 1982, xvi + 401 pp., $22.00, paperback.

Neuro-Physiologie des instincts et de la pensee (in Recherches Interdisciplinaires). Pierre Nelson. Maloine, Paris, 1982, 176 pp., paperback.

Energetics and Kinetics in Biotechnology

by **J.A. Roels**

The large scale use of organisms and parts thereof calls for quantitative understanding of the energetics and kinetics for biological transformation.
This present work develops the theory underlying the quantitative description of processes in which integral micro-organisms of enzymes are used. It clearly points out the relation to chemical engineering energetics and kinetics. In this book a number of useful, practical applications of the theory are treated and an extensive amount of engineering data are given. This volume will be invaluable to all researchers in microbial growth and product formation.

CONTENTS: Chapter 1. INTRODUCTION. Mathematical models. Kinetic models in biotechnology. The continuum theory. Conserved and non-conserved quantities. The first and second laws of thermodynamics. The kinetic description of transformation processes. *Chapter 2.* MACROSCOPIC THEORY FOR OPEN SYSTEMS. The mathematics of the macroscopic description. Conserved and non-conserved quantities. The balance equation. for the chemical state vector of a system. The balance equation for total mass. The elemental balance equations. The balance of energy and entropy. The linear phenomenological relationships. Conclusion. *Chapter 3.* MACROSCOPIC THEORY, AND MICROBIAL GROWTH AND PRODUCT FORMATION. Microbial systems as open systems. The elemental balance equation. Thermodynamic treatment of the energetics of growth. List of symbols. *Chapter 4.* THE LINEAR EQUATION FOR SUBSTRATE CONSUMPTION. The concept of maintenance energy. Aerobic growth without product formation. Anaerobic growth without external electron acceptors. Anaerobic formation of ethanol. The calculation of the anaerobic true yield and maintenance factor from its aerobic counterparts. *Chapter 5.* BIOCHEMICALLY STRUCTURED BALANCES OF MICROBIAL METABOLISM. Introduction. The concept of the ATP yield of growth. Aerobic growth, the P/O ratio. A biochemically structured model for aerobic growth on one substrate. A model for growth on mixed substrates. A model for growth with formation of a product. Conclusion. *Chapter 6.* THE KINETICS OF CHEMICAL REACTIONS. Introduction. Non-equilibrium thermodynamics of chemical reactions. The kinetics of chemical processes. The kinetic treatment of reaction patterns consisting of more than one reaction. Temperature dependence of the rate of chemical processes. *Chapter 7.* THE KINETICS OF ENZYME REACTIONS. Introduction. Kinetics of reactions involving one enzyme, one substrate and one product. Chains of consecutive enzyme reactions. Inhibition and activation phenomena in enzyme kinetics. Temperature effect in enzyme kinetics. *Chapter 8.* RELAXATION TIMES AND THEIR RELEVANCE TO THE CONSTRUCTION OF KINETIC MODELS. Mechanism and their relaxation times. Comparison of the relaxation times of mechanisms and the characteristic times of the observations. Interaction of the system's relaxation times with those of the environment. Relevance of the relaxation times concept to the modelling of bioengineering systems. *Chapter 9.* UNSTRUCTURED MODELS FOR GROWTH AND PRODUCT FORMATION IN MICROORGANISMS. Introduction. An unstructured model for biomass growth based on Monod's equation. Other kinetic equations for the substrate concentration dependence of the growth rate. Unstructured models for the formation of products.

1983 340 pages
Price: US$ 80.75
(in USA & Canada)
Dfl. 190.00 (Rest of World)
ISBN 0-444-80442-0

ELSEVIER BIOMEDICAL

Elsevier Science Publishers
P.O. Box 211, 1000 AE Amsterdam. The Netherlands
Distributor in the U.S.A. and Canada:
ELSEVIER SCIENCE PUBLISHING Co., Inc., 52 Vanderbilt Ave., New York, NY 10017

The Dutch guilder price is definitive. US $ prices are subject to exchange rate fluctuations.

BP 146

Information for Authors

Mathematical Biosciences publishes carefully selected mathematical papers of both research and expository type devoted to the formulation, analysis, and numerical solution of mathematical models in the biosciences.

Manuscripts should be submitted to the Editor: Dr. John A. Jacquez, Department of Physiology, University of Michigan, 7712 Medical Sciences II, Ann Arbor, Michigan 48109. Manuscripts are normally sent out for two independent reviews; usually a member of the Editorial Board and an outside reviewer serve as referees. However, outside referees will not necessarily be used for papers in the specialty areas of the members of the Editorial Board. Authors should indicate which members of the Editorial Board they recommend review their papers. Manuscripts are received with the understanding that their content is original work and is not being submitted for publication elsewhere. Authors should submit the original and two copies of the manuscript. The *copies* will be used in the reviewing process and will not be returned.

Upon acceptance of an article by the journal, the author(s) will be asked to transfer copyright in the article to the Publisher. This transfer will ensure the widest possible dissemination of information under the U. S. Copyright Law.

Form of manuscript. All material should be typed double spaced on $8\frac{1}{2} \times 11$-in. bond paper. Footnotes, reference lists, tabular material and figure legends should be typed on separate sheets. References to the literature are indicated in the text by full size Arabic numbers in brackets, i.e., [1]. Footnotes to the text are indicated by superior numbers. In the journal they will be placed at the foot of the page on which they are cited. An *Abstract* of about 200 words or less should be included. The *title page* should include full names of the authors, with the mailing address for proofs and correspondence. If the title of the article exceeds 45 characters and spaces, include a brief *running title*.

Mathematical notation. Use typewritten letters, numbers, and symbols wherever possible. Identify boldface, script letters, etc. the first time they occur. Distinguish between arabic "1" and the letter "l" and between zero and the letter "O", capital or lower case, wherever confusion might result.

The following *reference style* should be observed:

Journal. Gur Huberman, Qualitative behavior of a fishery system, *Math. Biosci.* 42:1–14 (1978).

Book. L. Nachbin, *Holomorphic Functions, Domains of Holomorphy, and Local Properties*, North-Holland/American Elsevier, Amsterdam, New York, 1972, p. 56.

Edited Book. (Author and chapter title, if desired, followed by *in*) *Spectral Theory and Asymptotics of Differential Equations*, E. M. de Jager, ed., North-Holland/American Elsevier, Amsterdam, New York, 1974, pp. 131–140.

Theses, Reports, and Other Unpublished Material. Style as a journal article with as much source information as possible.

Figures. Illustrations should be submitted as original line drawings or glossy photostats of originals, as professionally done as possible, complete and ready for photoreproduction. Lettering and data points should be large enough so when reduced to fit on the journal page ($4\frac{5}{8}$ in. width $\times$ 7 in. depth maximum) they can be read with ease.

Proofs and reprints. Authors receive proofs, which should be proofread, corrected and returned within 48 hours of receipt. Reprints may be ordered using the order form which accompanies proofs.

Call For Software

Elsevier Science Publishing Company, Inc. (New York) invites you to list your software packages, at no charge, in the
INTERNATIONAL SOFTWARE DATABASE™

Listings are disseminated online through DIALOG file 232 and in Elsevier's quarterly publications:

The Software Catalog: MICROCOMPUTERS
The Software Catalog: MINICOMPUTERS

Information required for a listing includes:

- Your name, address, and telephone number as a vendor or distributor
- Your terms of sale (e.g., personal check, C.O.D., credit cards) and acceptable currencies
- Names of the programs to be listed
- Whether a warranty, source code, or updates are provided for each program (and at what price, if any)
- Special hardware or software environments required
- Computers and operating systems on which each program runs
- Language in which each program is written
- Microprocessors for which they are written (if applicable)
- Minimum memory requirements
- Distribution medium
- Prices
- Subject areas and applications for which the programs are intended
- A succinct, factual description of the purposes and functions of each program

Each listed program is assigned a unique **INTERNATIONAL STANDARD PROGRAM NUMBER™ (ISPN™)** which serves as a cross-reference within **The Software Catalog** and as a package identifier for the software industry (including retailers, purchasers, distributors, and publishers).

For further information and forms, write to:
Elsevier Science Publishing Company, Inc.
Attn: Dr. Kenneth Metzner, Senior Editor
52 Vanderbilt Avenue
New York, NY 10017 USA

Elsevier